Title
A longitudinal big data approach for precision health.

Journal
Nature medicine, 25(5)

ISSN
1078-8956

Authors
Schüssler-Fiorenza Rose, Sophia Miryam
Contrepois, Kévin
Moneghetti, Kegan J
et al.

Publication Date
2019-05-01

DOI
10.1038/s41591-019-0414-6

Peer reviewed

HHS Public Access

Author manuscript

Nat Med. Author manuscript; available in PMC 2019 November 08.

Published in final edited form as:

Nat Med. 2019 May ; 25(5): 792–804. doi:10.1038/s41591-019-0414-6.

A Longitudinal Big Data Approach for Precision Health

Sophia Miryam Schüssler-Fiorenza Rose[1,2,3,*], **Kévin Contrepois**[1,*], **Kegan J Moneghetti**[4,5,6], **Wenyu Zhou**[1], **Tejaswini Mishra**[1], **Samson Mataraso**[7,8], **Orit Dagan-Rosenfeld**[1], **Ariel B. Ganz**[1], **Jessilyn Dunn**[1,9], **Daniel Hornburg**[1], **Shannon Rego**[1], **Dalia Perelman**[1], **Sara Ahadi**[1], **M. Reza Sailani**[1], **Yanjiao Zhou**[10,11], **Shana R. Leopold**[10], **Jieming Chen**[12], **Melanie Ashland**[1], **Jeffrey W Christle**[4,5], **Monika Avina**[1], **Pats Limcaoco**[1], **Camilo Ruiz**[13], **Marilyn Tan**[14], **Atul J Butte**[12], **George M Weinstock**[10], **George M. Slavich**[15], **Erica Sodergren**[10], **Tracey L. McLaughlin**[14], **Francois Haddad**[4,5,**], **Michael P Snyder**[1,4,**]

[1]Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305, USA

[2]Spinal Cord Injury Service, Veteran Affairs Palo Alto Health Care System, Palo Alto, CA 94304, USA

[3]Department of Neurosurgery, Stanford University School of Medicine, Stanford, CA 94305, USA

[4]Stanford Cardiovascular Institute, Stanford University, Stanford, CA, 94305 USA.

[**]To whom correspondence should be addressed: mpsnyder@stanford.edu or fhaddad@stanford.edu.

Author Contributions

S.M.S.-F.R., M.P.S., F.H., K.C., K.M., T.M., W.Z. contributed to conceptualization. S.M.S.-F.R., K.C., F.H., M.P.S., T.M., K.M., S.M., W.Z., S.R. contributed to methodology. K.C. (ASCVD biomarkers), D.H. (Lipidomics), A.B.G. (Microbiome DADA2 processing), T.M., M.A., W.Z. (OGTT c-peptide and insulin) contributed to omics generation and/or processing. S.M.S.-F.R., K.C., T.M., W.Z., J.D., M.A., J.W.C., E.S., P.L. contributed to data curation. K.C., S.M.S.-F.R., T.M., K.M., F.H., M.P.S. contributed to visualization. S.M.S.-F.R., K.C., T.M., S.M., K.M., O.D.-R., S.R., J.C., C.R. contributed to formal analysis. S.M.S.-F.R., K.C., M.P.S. contributed to project administration. M.P.S., F.H. contributed to supervision. S.M.S.-F.R., F.H., K.C., K.M., M.P.S. contributed to writing and preparing the original draft. S.M.S.-F.R., K.C., K.M., F.H., M.P.S., W.Z., A.B.G., D.H., J.D., G.M.S, T.M., M.T., D.P., T.L.M., A.J.B., M.R.S., S.A. contributed to review and editing. K.M., F.H., J.W.C. contributed to cardiovascular clinical data collection and investigation. W.Z., S.R., M.A., P.L., D.P., M.T., T.L.M., S.M.S.-F.R. contributed to iPOP/iHMP clinical data collection/investigation. W.Z., S.R.L, M.P.S., T.L.M., E.S., G.M.W. contributed to iPOP/iHMP project administration. K.C. (metabolomics), S.A. (proteomics), M.R.S. (DNA, RNA-seq), W.Z. (microbiome, cytokines, and overall omics data), Y.Z. (microbiome), T.M. & D.H. (batch correction methodology for proteomics) contributed to iPOP/iHMP omics raw data processing. M.P.S., G.M.W., T.L.M., E.S. contributed to iPOP/iHMP funding acquisition.

[*]These authors contributed equally to this work

M.P.S. is a cofounder of Personalis, SensOmics, January, Filtricine, Qbio and Akna and an inventor on provisional patent number 62/814,746 'Methods for evaluation and treatment of glycemic dysregulation and applications thereof '. S.M.S.-F.R., K.C., W.Z., T.M. and S.M. are also listed as inventors. A.J.B. reports grants and non-financial support from Progenity, grants and personal fees from NIH (multiple institutes) and Genentech, and grants from L'Oreal, personal fees from NuMedii, Personalis, Lilly, Assay Depot, Geisinger Health, GNS Healthcare, uBiome, Roche, Wilson Sonsini Goodrich & Rosati, Orrick, Herrington & Sutcliffe, Verinata, 10x Genomics, Pathway Genomics, Guardant Health, Gerson Lehrman Group, Nuna Health, Samsung, Capital Royalty Group, Optum Labs, Pfizer, AbbVie, Bayer, Three Lakes Partners, HudsonAlpha, Tensegrity, Westat, FH Foundation, WuXi, FlareCapital, Helix, Roam Insights, Autodesk, Regenstrief Institute, American Medical Association, Precision Medicine World Conference, and Mars during the conduct of the study. A.J.B. has pending patent Atul J. Butte, Keiichi Kodama, Methods for diagnosis and treatment of non-insulin dependent diabetes mellitus, published August 4, 2011, WO2011094731 and US20130071408; patent Joel T. Dudley, Atul J. Butte, Method and System for Computing and Integrating Genetic and Environmental Health Risks for a Personal Genome, published April 26, 2012, US20120101736 with royalties paid to Personalis; patent Joel T. Dudley, Atul J. Butte, Method And System For Functional Evolutionary Assessment Of Genetic Variants, published April 11, 2013, US20130090909 with royalties paid to Personalis; patent Konrad Karczewski, Michael Snyder, Atul J. Butte, Joel T. Dudley, Eurie Hong, Alan Boyle, J. Michael Cherry, Method and System for Assessment of Regulatory Variants in a Genome, published May 9, 2013, US20130116931 with royalties paid to Personalis; and patent Frederick Dewey, Euan Ashley, Carlos Daniel Bustamante, Atul Butte, Jake Byrnes, Rong Chen, Phased Whole Genome Genetic Risk In A Family Quartet, published March 28, 2013, US20130080068, with royalties paid to Personalis; Stanford University pays royalties each year on licensed intellectual property.

[5]Division of Cardiovascular Medicine, Department of Medicine, Stanford University School of Medicine, Stanford, CA, 94305 USA

[6]Department of Medicine, St Vincent's Hospital, University of Melbourne, Melbourne, Australia

[7]Department of Electrical Engineering and Computer Sciences, University of California - Berkeley, Berkeley, CA 94720

[8]Department of Bioengineering, University of California - Berkeley, Berkeley, CA 94720

[9]Mobilize Center, Stanford University, Stanford, California, USA, 94305

[10]The Jackson Laboratory for Genomic Medicine, Farmington, CT 06032, USA

[11]Department of Medicine, University of Connecticut Health, Farmington, CT 06030, USA

[12]Bakar Computational Health Sciences Institute and Department of Pediatrics, University of California, San Francisco, California 94143, USA

[13]Department of Bioengineering, Stanford University, Stanford CA, 94305, USA

[14]Division of Endocrinology, Stanford University School of Medicine, Stanford, CA 94305, USA

[15]Cousins Center for Psychoneuroimmunology and Department of Psychiatry and Biobehavioral Sciences, University of California, Los Angeles, CA 90095, USA

Abstract

Precision health relies on the ability to assess disease risk at an individual level, detect early preclinical conditions and initiate preventive strategies. Recent technological advances in omics and wearable monitoring enable deep molecular and physiological profiling and may provide important tools for precision health. We explored the ability of deep longitudinal profiling to make health-related discoveries, identify clinically relevant molecular pathways, and impact behavior in a prospective longitudinal cohort (n = 109) enriched for risk of type 2 diabetes mellitus (DM). The cohort underwent integrative Personalized Omics Profiling (iPOP) from samples collected quarterly for up to 8 years (median 2.8 years) using clinical measures and emerging technologies including genome, immunome, transcriptome, proteome, metabolome, microbiome, and wearable monitoring. We discovered over 67 clinically actionable health discoveries and identified multiple molecular pathways associated with metabolic, cardiovascular and oncologic pathophysiology. We developed prediction models for insulin resistance using omics measurements illustrating their potential to replace burdensome tests. Finally, study participation lead the majority of participants to implement diet and exercise changes. Altogether, we conclude that deep longitudinal profiling can lead to actionable health discoveries and provide relevant information for precision health.

Introduction

Precision health and medicine are entering a new era where wearable sensors, omics technologies, and computational methods have the potential to improve health and lead to mechanistic discoveries[1,2]. Emerging technologies such as longitudinal multi-omics profiling combined with clinical measures can comprehensively assess health and identify deviations from healthy baselines which may improve disease risk prediction and early

detection. Connecting longitudinal multi-omics profiling with clinical assessment is also important in developing a new taxonomy of disease based on molecular measures[1].

Despite this promise, few studies have leveraged emerging technologies and longitudinal profiling to manage health and identify disease markers. Previous efforts included our study of a single individual in which longitudinal multi-omics profiling over 14 months captured the individual's transition to diabetes on a deep molecular level[3]. A recent study of 108 individuals followed for 9 months using various omic technologies revealed several health-related findings[4]. A cross-sectional study used genome sequencing, metabolomics and advanced imaging to identify individuals at risk for age-related chronic disease[5]. These studies either had limited sample size, lacked meaningful longitudinal profiling, or performed only limited analysis of health information. We have also demonstrated utility in using wearable devices to detect infections[2] and identify early glucose dysregulation[6] and population-based studies are underway to potentially to detect arrhythmias[7].

In this study, we longitudinally profiled 109 participants at risk for DM (Fig. 1), performing quarterly clinical laboratory tests and multi-omics assessments. In addition, individuals underwent exercise testing, enhanced cardiovascular imaging and physiological testing, wearable sensor monitoring, and completed various surveys.

The study objectives were threefold. We first evaluated the usefulness of emerging technologies in combination with standard and enhanced clinical tests to detect diseases early. We then characterized multi-omics associations with clinical pathophysiologies including glucose and insulin dysregulation, inflammation, and cardiovascular risk; evaluated the ability of multi-omics measures to predict insulin resistance and response to glucose load. Lastly, we examined how participation affected health habits.

Results

Summary of Research Design & Cohort

A 109-person cohort enriched for individuals at risk for DM (**Table 1**, Extended Data Fig. 1a) underwent quarterly longitudinal profiling for up to eight years (median 2.8 years) using standard and enhanced clinical measures and emerging assays. (Fig. 1). Emerging tests included molecular profiling of the genome, gene expression (transcriptome), proteins (proteome), immune proteins (immunome), small molecules (metabolome) and gut microbes (microbiome), and wearable monitoring including continuous glucose monitoring (CGM)[6]. Our study was designed to capture transitions from normoglycemic to preDM and from preDM to DM. Thus, in addition to standard measures such as fasting plasma glucose (FPG, reflects steady state glucose metabolism[8]) and glycated hemoglobin (HbA1C, reflects 3 month average glucose), enhanced measures included the oral glucose tolerance test (OGTT, reflects response to glucose load[9]) with insulin secretion assessment (beta-cell function) and the modified insulin suppression test (SSPG, a measure of peripheral insulin resistance). We also performed enhanced cardiovascular profiling including vascular ultrasound, echocardiography, cardiopulmonary exercise testing and cardiovascular disease protein markers. Technical details are provided in the methods and our integrated Human Microbiome Project (iHMP) paper by Zhou et al. (submitted). The full details of clinical

laboratory measures, immune proteins and cardiovascular biomarkers are provided in Table S0. The study was approved by the Stanford University Institutional Review Board (IRB 23602) and all participants consented.

The mean age of iPOP participants at initial enrollment was 53.4 ± 9.2 years old. Demographic, baseline health, and family history characteristics are shown in Table S1. Genetic ancestry analysis (n = 72) using the 1000 Genomes data[10] shows that individuals mapped to expected ancestral populations (Extended Data Fig. 1b).

Over the study course, we found over 67 major clinically actionable health discoveries spanning metabolism, cardiovascular disease, oncology and hematology, and infectious disease (Table S2). We demonstrate ways in which longitudinal multi-omics measures can be used to advance precision medicine, including by illuminating biological pathways underlying standard measures, predicting burdensome physiological measurements, and enabling exploration of mechanisms of disease onset.

Metabolic Health Profiling

At entry, participants reported their DM status. Of the 86 participants (78.9%) who did not report preDM or DM, one had a diagnosis of DM in their health record, one had a DM-range HbA1C and 43 individuals (39.4%) had labs in the preDM range at entry (Fig. 2a). During the study, eight more individuals converted to DM as assessed by a clinical diagnosis of DM (n = 4), starting a diabetic medication after a diabetic range laboratory result (n = 3), and/or if they had labs in the diabetic range (n = 6) at more than one time point. Five additional participants developed laboratory abnormalities in the diabetic range at one time point, and 12 developed abnormalities in the prediabetic range. In addition, 2 participants had diabetic range CGM measurements (> 200 mg/dL) who were normoglycemic on FPG, HbA1C and OGTT (Table S3) indicating that these individuals have glucose dysregulation that is most easily assessed using CGM.

Value of exome sequencing—Exome sequencing[11] provided relevant information for diabetes management. Most notable was the discovery of a hepatic nuclear factor 1A mutation, pathogenic for Maturity-Onset Diabetes of the Young (MODY), in a participant with DM. This discovery has implications for medications[12] and the individual decided to have the children tested. Excluding a MODY mutation was valuable to a second participant. Other discoveries are listed in Table S2.

Enhanced metabolic profiling—DM is a complex disease with various underlying pathophysiologies including insulin resistance, pancreatic beta-cell dysfunction and abnormal gluconeogenesis[13], which can have differential effects on standard measures. Over the study course, 22 participants had at least one test result in the diabetic range (Fig. 2b) but few (n = 2) had concordance of all three measures. When performed simultaneously, FPG-HbA1C and FPG-OGTT were in agreement 65.2% and 52.6% of the time, respectively (Extended Data Fig. 2a,b), highlighting that DM status varies depending on the assessment method. Most participants also underwent insulin sensitivity assessment (n = 69); 55% were resistant (SSPG ≥ 150 mg/dl). In addition, insulin secretion during OGTT was assessed in 61 participants using the C-peptide deconvolution method[14] and the glucose disposition

index (DI) was calculated[15]. Based on OGTT measurements, participants were categorized in three groups: normoglycemic, impaired fasting glucose only (IFG only) and impaired glucose tolerance (IGT). We observed large inter-individual variability in insulin levels, insulin resistance and DI between groups (Fig. 2c). Participants with IGT had higher insulin levels 120 min post-OGTT test, higher SSPG (more insulin resistant) and a lower DI. Cluster analysis of the longitudinal pattern of insulin secretion rates during OGTTs demonstrated four insulin secretion groups: early, intermediate, late and very late (Fig. 2d). Each cluster was heterogeneous in terms of OGTT status, DI, insulin resistance status and maximum insulin level and demonstrated no consistent pattern of molecular enrichment, indicating high heterogeneity in glucose dysregulation.

We also searched for multi-omics molecular associations with the disposition index across the cohort and found 109 significant molecules (FDR < 0.1) (Table S4). HbA1C (FDR = 2.0E-03) and FPG (FDR = 4.9E-02) were negatively associated with DI as expected from previous reports showing increased FPG and HbA1C with beta-cell dysfunction[16,17]. We found that DI was strongly negatively associated with leptin (FDR=1.6E-07) and GM-CSF (FDR=7.2E-07) which are known regulators of energy homeostasis and inflammation signaling[18,19]. GM-CSF (p = 1.5E-07) and leptin (p = 3.3E-07) were also the two analytes that were most strongly positively associated with body mass index in our cohort and were positively associated with hsCRP illustrating their connection to inflammation and obesity. In the DI correlation network, leptin and GM-CSF were correlated with various lipid classes including an inverse correlation with androgenic steroids, and a positive correlation with sphingolipids and sphingosines, free fatty acids and glycerophospholipids highlighting their importance in lipid metabolism[20] (Fig. 2e, Table S5).

Longitudinal course & mechanistic insights—A study strength is its dense longitudinal sampling approximately every 3 months. Based on individual longitudinal HbA1C trajectories, participants were classified into 6 categories (Extended Data Fig. 2c). Notably it was common for participants' HbA1C to alternate between normal-preDM (n = 21) and preDM-DM range (n = 8). No one stayed exclusively within the DM range due to good diabetes control with lifestyle and medications. Consistent transitions from normal to preDM (n = 5) and from preDM to normal HbA1C (n = 10) were less common.

Close evaluation of individual trajectories of participants with new diabetes (n = 9) revealed additional insights. Individual trajectory analysis revealed that participants followed multiple pathways to diabetes (Fig. 3a-c, Extended Data Fig. 3, Table S3). Some participants' (n = 2) first abnormality was DM-range OGTT (Fig 3a, Extended Data Fig. 3a), others (n = 3) had elevated FPG (Fig. 3b, Extended Data Fig. 3b,c), the remainder (n = 4) had a DM-range HbA1C (Extended Data Fig. 3d,e) or abnormalities in multiple measures (Fig. 3c, Extended Data Fig. 3f). Interestingly, diabetic range labs followed viral infections[3] in one participant (Fig. 3c). Also, one participant with a single DM lab improved their SSPG with diet and exercise (Extended Data Fig. 3g) and never had a second DM range lab during the study.

Progression to DM was associated with weight gain and decreased gut microbiome diversity (Shannon) in 2 of 8 participants (Fig. 3a,b, Extended Data Fig. 4a,b). In both cases, the phylum Bacteroidetes proportion was increased at the time point of lowest diversity to the

detriment of beneficial bacteria such as the genus faecalibacterium (Extended Data Fig. 4c,d,e). Using linear mixed models to account for repeated measures, we evaluated the relationship between microbiome diversity and SSPG, FPG and HbA1C and found an inverse relationship with diversity that was strongest with SSPG (p = 1.5E-04) (Table S6). We then performed longitudinal mixed model analysis to understand changes in diversity over time (Table S7). SSPG accounted for 28% of the between-person Shannon variance highlighting the importance of insulin resistance in microbiome diversity. The majority of Shannon variance was intra-individual (76.8%) and adding the Bacteroidetes phylum proportion to the model including its interaction with time accounted for 41% of the remaining within-person variance, consistent with the relationship observed in the individual profiles between Bacteroidetes proportion and diversity.

Longitudinal evaluation of all data related to glucose and insulin regulation provided insights into mechanism. For instance, the person in Fig. 3c had a normal SSPG despite a diabetic range OGTT, FPG and HbA1c. Although elevated OGTT is commonly thought to result from increased peripheral resistance or decreased insulin production, this participant had elevated insulin production with a delayed response trajectory, possibly reflecting delayed insulin release (Table S3). Other mechanistic insights are provided in Table S3. In conclusion, participants developed diabetes through different pathways and our detailed characterization provides potential hypotheses regarding individual underlying mechanism of glucose dysregulation which is a goal of precision medicine.

Multi-omic dimensions of glucose metabolism & inflammation—We examined the underlying relationships between glucose (FPG, HbA1C) and inflammation (hsCRP) levels and multi-omics measurements at healthy time points (healthy-baseline models) and with relative changes for all time points (dynamic models) using linear mixed models. The two analyses are complementary since the healthy-baseline models highlight the stable relationships between measures and dynamic models highlight common associations with change.

As expected, the healthy-baseline analysis demonstrated that HbA1C and FPG strongly associated with each other and the ‘glucose homeostasis’ pathway (Fig. 3d, Extended Data Fig. 5, Tables S8-13). Although the two measures had many common associations, particularly with metabolites including lipids (free fatty acids and total triglyceride level (TGL)) and amino acids as previously reported[21], many analytes were exclusively associated with FPG or HbA1C highlighting the differential underlying biology captured by both measures. While HbA1C associated with unsaturated fatty acid (FDR = 8.2E-04) and glycerophospholipid metabolism (FDR = 2.88E-03), FPG associated with amino acid (FDR = 7.4E-04) and bile acid metabolism (FDR = 4.6E-03).

The dynamic model analysis revealed more commonalities between changes in glucose measures and inflammation (Fig. 3d,e, Extended Data Fig. 5, Tables S14-19). As expected, hsCRP positively associated with inflammatory proteins including MIG (FDR = 1.4E-24) and IP10 (FDR = 3.9E-22) as well as immune pathways including ‘complement activation’ (FDR = 8.7E-16), ‘innate immune system’ (FDR = 8.3E-14) and ‘oxidative damage’ (FDR = 3.0E-06). Interestingly, both HbA1C and hsCRP positively associated with total white blood

cells, monocytes and neutrophils consistent with previous findings[22]. In addition, hepatocyte growth factor (HGF) associated with HbA1C and hsCRP, consistent with its role in glucose metabolism and modulation of inflammatory response[23]. We also observed that FPG and HbA1C both associated with 'leukotriene biosynthesis' which contributes to inflammation and leads to insulin resistance[24]. HbA1C also associated with additional pathways related to lipid metabolism including 'plasma lipoprotein assembly' and 'chylomicron assembly' which further demonstrates the connection between inflammation, lipid metabolism and metabolic regulation of glucose.

Multi-omics prediction of SSPG & OGTT—The modified insulin suppression test is a clinically important direct measure of peripheral insulin resistance but is expensive, labor-intensive, and requires six hours. The two-hour OGTT is a sensitive test for diabetes and is less expensive, but still inconvenient. Thus, we evaluated how well multi-omics measurements could predict the results of these tests. Using a Bayesian network algorithm, we first identified highly predictive features followed by ridge regression modeling using these features[25,26]. The SSPG prediction model using all omes achieved a cross-validated R^2 of 0.87 (final model mean square error (MSE) 0.16) compared to an R^2 of 0.59 (MSE 0.55) using clinical data only (Fig. 3f, Table S20). We also compared predictive models using clinical data plus each single ome and found that the transcriptome (R^2 0.88, MSE 0.15), metabolome (R^2 0.80, MSE 0.31) and microbiome models (R^2 0.78, MSE 0.26) had the best predictive accuracy for SSPG. Similarly, the multi-omic prediction model for OGTT (R^2 0.71, MSE 0.24) was superior to the clinical data only model (R^2 0.42, MSE 0.71) (Fig. 3f, Table S21). The transcriptome had the best predictive accuracy of the single ome models (R^2 0.62, MSE 0.30). Molecules that were found to be consistent across multiple SSPG models included the TGL/HDL (high-density lipoprotein) ratio, the protein IL-1RAP; the lipid Hexosylceramide (HCER)(24:0), the MAP3K19 transcript and a Ruminococcaceae family microbe. The relationship between insulin resistance and TGL/HDL ratio has already been described[27] and other measures are emerging[28-30]. There was little overlap between SSPG and OGTT predictors supporting that these measures reflect different underlying biology. The increased predictive performance with multi-omics measurements compared to clinical labs alone illustrates the benefit of multi-omics data.

Other metabolic disorders—Other clinical abnormalities were observed in sodium, potassium and liver enzymes (ALT) as well as microalbuminuria and macroalbuminuria (Table S2). People with preDM and DM are at higher risk for liver steatosis and albuminuria. Using the American Gastroenterological Association (AGA) Guidelines [31] for health normal references (males: 25–33 IU/L; females: 19–25 IU/L) revealed that the majority of participants (83%) had at least one elevated healthy visit ALT and 41% had elevations at all healthy time points. Given the AGA recommendations for ultrasound screening[31], our findings suggest that screening for nonalcoholic fatty liver disease is indicated in the majority of our population.

One participant was a significant outlier in gene expression related to toxicity pathways including oxidative stress and hepatic abnormality pathways (Extended Data Fig. 6a, Zhou et al., submitted). The participant had mild elevation in ALT accompanied by increases in

bile acids and glutamyl dipeptides (Extended Data Fig. 6b), and was later diagnosed with mild hepatic steatosis. However, many participants had mild ALT elevations and at least five had hepatic steatosis, thus these clinical findings are not sufficient to explain the RNA-seq outlier status. Although multiple omics and other measures point to aberrant hepatic function, clinical manifestations were unclear and this individual will be tracked for hepatic abnormalities.

Cardiovascular Health Profiling

Atherosclerotic cardiovascular disease (ASCVD) is a major cause of mortality and morbidity associated with insulin resistance and DM[32]. We assessed the American Heart Association ASCVD risk score, estimating 10-year risk of heart disease or stroke on all participants[33] at study entry. We also followed longitudinal trajectories of dyslipidemia and systemic hypertension. Enhanced cardiovascular profiling was performed on 43 participants and included i) vascular ultrasound and echocardiography to assess for subclinical atherosclerosis, arterial stiffness or early stage adverse ventricular remodeling or dysfunction and ii) emerging biomarkers assessment to interrogate oxidative stress, inflammation, immune regulation, myocardial injury and myocardial stress pathways[34-36].

Cardiovascular risk profiles—At study entry, 24 patients (22.6%) had an ASCVD risk score ≥ 7.5%, a threshold often used to guide primary prevention[33] (Fig. 4a). Total cholesterol and blood pressure measurements indicate that self-report underestimated the prevalence of dyslipidemia (Fig. 4b) and 18 participants learned they had Stage II hypertension during the study.

Clinical discoveries through enhanced clinical phenotyping—Wearable and cardiovascular imaging led to important clinical discoveries. Wearable heart rate monitoring identified two participants with nocturnal supraventricular tachycardia, leading to the diagnosis of obstructive sleep apnea in one and atrial fibrillation secondary to sleep apnea in the other. In the subgroup of participants who had enhanced cardiovascular imaging studies, we discovered two major health findings: one cardiac finding associated with a pathogenic mutation in the RPM20 gene, and one non-cardiac finding (Table S2). Fitness assessment using percent predicted oxygen consumption (maximal oxygen consumption relative to a healthy person of the same age and weight) identified three participants with values below 70% suggestive of a reduction in exercise capacity which has been associated with poorer health outcomes[37] (Extended Data Fig. 7a). Subclinical atherosclerosis was found in six participants leading to a recommendation to increase statin dose (Extended Data Fig. 7b). Overall, there were 15 important clinical findings through these enhanced tests (Table S2).

Cardiovascular events, pharmacogenomic & transcriptomic findings—Five participants had cardiovascular events during the course of the study including stroke (n = 3), unstable angina (n = 1) and stress-induced cardiomyopathy (n = 1). All had elevated hsCRP levels prior to their event. Two participants with incident strokes had pharmacogenomic variants that could partially explain suboptimal response to the chosen therapy. One participant on aspirin for stroke prevention had a COMT (catechol-o-methyltransferase) Val/Val genotype (rs4680) which has a 85% increased risk of

cardiovascular events in female aspirin users compared to placebo controls[38]. The other participant with incident stroke was an intermediate clopidogrel metabolizer phenotype (CYP2C19*2 (rs4244285)/CYP2C19*17 (rs12248650) and had a second stroke while on clopidogrel. Intermediate metabolizers of clopidogrel were common in our study (31/88 (35%)) and 4/88 (4.5%) were poor metabolizers. Additional pharmacogenomic variants related to the common cardiovascular medications statins and coumadin were found in 26 and 30 participants, respectively (Table S22).

We also analyzed 14 of 32 genes associated with stroke and stroke types[39] which were robustly detected in our RNA-seq dataset. Outlier analysis revealed that two of the five participants with cardiovascular events had the highest composite Z-scores at clinically relevant time points including post-stent placement (Z-score = 33.2, FDR = 6.9E-06), mid-infection (Z-score = 40.4, FDR = 3.2E-09) for one participant and transition to diabetes (Z-score = 30.1 and 24.1) for the other (Extended Data Fig. 7d,e). Thus, expression levels of genes related to stroke were outliers and associated with significant health issues.

Multi-omics analysis of ASCVD risk—We evaluated multi-omics measures associated with adjusted ASCVD risk score using Spearman correlation (Table S23), and constructed a correlation network. This analysis revealed relationships between clinical and omics measures such as monocytes bridging cytokines and complement proteins, and triglyceride and cholesterol measures linking to apolipoproteins (Fig. 4c, Table S24). Among immune proteins, the interferon-gamma pathway (MIG, IP10, interleukin (IL)-2, vascular endothelial growth factor alpha and HGF) were strongly associated with the ASCVD risk score. The interferon-gamma pathway has been recently found to play a key role in atherosclerosis based on population based studies[40-44]. IL-2 has been shown to be associated with atherosclerosis through its role in T-cell mediated inflammation[44]. HGF is involved in the survival of endothelial cells and is emerging as a risk factor of outcome[41,42]. Our network also highlighted several molecules that are emerging in cardiovascular disease including complement and free fatty acids as well as γ-glutamyl-ε-lysine (reported in diabetic nephropathy), hypoxanthine, methylxanthine (associated with coffee consumption) and bile acids[45-47].

In participants who underwent cardiovascular imaging, we also performed a correlation network analysis that shows how ASCVD risk, enhanced imaging and selected circulating protein markers associate together (Extended Data Fig. 7c, Table S1). ASCVD score was closely related to HGF, which itself was closely related to inflammatory cytokines IL-1B and IL-18, part of the inflammasome complex. Exercise capacity as assessed with peak VO_2 was closely associated with GDF-15, a transforming growth factor which is associated with cardiovascular mortality risk[48] and leptin, a hormone that regulates appetite[49]. These findings demonstrate an interaction between inflammation and ASCVD risk and suggest new opportunities for personalized risk stratification, beyond those currently available.

Oncological, Hematological & Immune Profiling

Exome sequencing also led to several important oncological, hematological and immune-related clinical discoveries. Eight participants learned they had clinically actionable genetic

variants associated with increased oncologic risk, such as APC, SDHB, BRCA1, MUTYH, CHEK2 and hematologic risk (PROS1) (Table S2). In one case, follow-up screening led to discovery of an early-stage papillary thyroid cancer, and the participant was able to elect thyroid preserving surgery due to early detection.

B-cell lymphoma discovery & longitudinal outlier analysis—Abdominal ultrasound imaging revealed splenomegaly and large para-aortic lymph nodes in one participant (Fig. 5a); immediate clinical work-up (Fig. 5b,c, Table S3) led to diagnosis of B-cell lymphoma. Longitudinal omics outlier analysis revealed a striking increase (> 5-fold) in the cytokine MIG that started over a year prior to diagnosis and returned to baseline after treatment (Fig. 5d). Its early elevation suggests possible utility as an early biomarker, consistent with other studies[50-52]. Although likely important in a number of cancers[53], our data demonstrates MIG's utility as a longitudinal marker of disease. A notable decrease in histidine-rich glycoprotein was also evident at diagnosis (Table S25), consistent with its previously reported role in inhibiting tumor growth and metastasis[54,55].

The functional association network using proteins which were in the 95th percentile at the time of diagnosis relative to all the healthy visits in the study illustrates the central role of MIG in orchestrating other cytokines, namely ENA78, IL17A and VCAM1 (Fig. 5e). Pathways involved in inflammation/immune response as well as cell proliferation and migration were enriched at time of diagnosis (Table S26). The participant's gut microbiome Shannon diversity also changed with time ($p = 0.0041$), primarily declining in the two years prior to diagnosis, with a nadir at diagnosis (Fig. 5f) and increasing with treatment. Outlier microbes (95 percentile) at time of diagnosis included low proportions of the genera Clostridium IV, Lachnospiraceae incertae sedis, unclassified Clostridiales and Ruminococcaceae and elevated proportions of the class Gammaproteobacteria (Table S25). Similar to our findings in participants with low diversity prior to DM diagnosis, at the point of lowest diversity, the phylum Bacteroides predominated (84%). Altogether, we demonstrate that longitudinal molecular outlier analysis can identify deviations in key molecules associated with disease to reveal potential biomarkers and give insights into underlying biological mechanisms associated with the disease.

Hematologic, immune & infection profiling—Comprehensive clinical labs identified many important health-related findings. Thirty participants had hemoglobin or hematocrit in the anemic range, including 28 participants without prior known anemia (Hemoglobin: Males <13.5 g/dL, Females <11.7 g/dL). In participants with anemia, mean corpuscular volume (MCV) was low (< 82 f/L) in 26.7% ($n = 8$) suggesting microcytic anemia, 10% ($n = 3$) had an elevated MCV (> 98 f/L) with normal mean corpuscular hemoglobin concentration and the remainder had normocytic anemia. Importantly, one of these participants was discovered to have alpha thalassemia trait after referral to their physician for anemia evaluation.

Immunological profiling with IgM, identified one participant with a significantly elevated IgM (Fig. 5g) which led to a clinical diagnosis of monoclonal gammopathy of undetermined significance (MGUS). Nine participants were noted to have persistently low IgM (2 or more IgM < 30 mg/dL). Four participants had subsequent clinical evaluation of IgA and IgG

which led to identification of IgG monoclonal gammopathy and subsequent diagnosis of smoldering myeloma in one participant. The discovery of MGUS and smoldering myeloma precancers has important implications in elevated risk and screening for cancer[56,57].

During the study, wearable monitoring detected temperature and heart rate abnormalities related to inflammatory disturbances as measured by hsCRP ($n = 4$). In one of the participants, these findings resulted in diagnosis of Lyme disease[2]. Thus, important health information related to hematologic, immune and infection systems were revealed by a variety of different approaches.

Effect of iPOP Participation on Participants

The deep phenotyping profiling had an effect on the majority of the participants by (a) encouraging appropriate risk-based screening including genetic counseling, (b) facilitating clinically meaningful diagnosis, (c) potentially informing therapeutic choices (mechanistic or pharmacogenomic information), and (d) increasing awareness leading to diet and physical activity modifications. Overall, we found over 67 major clinically actionable health discoveries spanning various area including metabolic, cardiovascular, heme/oncological and infectious using standard clinical, enhanced, and emerging technologies (Fig. 6a, Table S2).

Fifty-eight participants were surveyed mid to late study about the effect of participating in the study including changes on food and exercise habits, health findings, and their sharing of results with their personal doctors, family and others. Eighty-two percent reported some change in diet and/or exercise habits (Fig. 6b). In addition, almost half reported changing other health behaviors as a result of the study, including improving sleep, reducing stress, adding fiber and supplements to their diet, more careful self-examinations, recording food intake, attending a fitness camp and general lifestyle changes (Table S27). Fig. 6c shows the amount of change in diet and exercise. Participants also reported that their wearable device kept them accountable for exercising and more mindful to take walking breaks. Others reported using wearables to monitor sleep.

The majority of participants had discussed study results with their family (71%) and physicians (68%). Physician discussions led to follow-up testing in 29% of the cases. Additional testing included having children tested for gene mutation, colonoscopy, additional eye exams, cardiac calcium scan, PET scan to evaluate lymphoma, repeating study tests (echocardiogram, pulmonary function tests) in the clinical setting, extra screening for macular degeneration risk, and additional tests for diabetes-related studies (SSPG and the Quantitative Sudomotor Axon Reflex Test). Participants were also asked about the effect of SSPG testing and CGM monitoring (Table S28). Eight participants who used a CGM monitor reported that it helped them make different dietary and meal frequency choices to reduce their blood sugar spikes. SSPG results motivated at least 2 participants to change their activity and diet and were reassuring to others. Therefore, overall, a myriad of positive behavior modifications and follow-up tests resulted from study participation.

Discussion

Our study found that combining untargeted multi-omics and physiological longitudinal profiling with targeted profiling of metabolic and cardiovascular risk led to actionable health discoveries and meaningful physiological insights building on our previous work[3]. Our targeted profiling approach enabled us to connect longitudinal profiling of glucose metabolism with multi-omics profiling facilitating the precision medicine goal of defining diseases based on molecular mechanisms and pathophysiology[1]. The untargeted longitudinal big data approach led to a number of discoveries in other areas such as cardiology, oncology, hematology and infectious disease, indicating that broad profiling is valuable for disease detection in many different areas. We capitalized on the depth of longitudinal profiling to identify deregulated molecules and pathways associated with the transition from health to disease.

The study informed more than half the participants of their preDM and DM status, dyslipidemia, and hypertension, which led many to institute diet and physical activity lifestyle changes. Our enhanced clinical assays including OGTT, beta-cell function assessment, insulin resistance and CGM in combination with standard clinical tests (FPG and HbA1C) improved characterization of preDM and DM status. Importantly, the in-depth physiological profiling identified individual mechanisms of glucose dysregulation which has important implications for implementation of personalized treatments. Our findings are consistent with the recent study which found that treatments based on the current classification are not well tailored to mechanistic subtypes[58] and proposed 5 subtypes of adult onset DM. Deeper molecular understanding of progression to DM and its characteristics in the individual may help tailor therapy to its underlying pathophysiology and will likely identify additional subtypes and also inform stratification of CVD risk[59]. The superiority of using multi-omics data for SSPG prediction compared to standard measures illustrates the value of multi-omics data to help provide a molecular taxonomy of disease[1], as well as replace expensive burdensome tests for insulin resistance with a simple blood test. Microbiome measures were also a good predictor of SSPG when combined with clinical measures and SSPG inversely correlated with Shannon diversity further demonstrating the intricate relationship between gut microbes and insulin resistance consistent with our multi-omics study of weight gain[60].

Although the majority of our exome sequencing findings were in the oncologic realm, several important metabolic exome findings were found including a MODY mutation with implications for medication management, a RBM20 mutation related to dilated cardiomyopathy and numerous pharmacogenomic variants that have important health implications[61]. Furthermore, two participants experienced vascular events, unaware of relevant pharmacogenomics information which could have suggested alternative treatments. Thus, we expect complex genetic risk assessment such as the information learned in this study to be incorporated into risk management and tailored treatment of disease[62].

Imaging plays a central part in precision health initiatives allowing the early detection of oncological and systemic disease[63]. In our study, imaging helped detect dilated cardiomyopathy (in the RBM20 patient), early-stage atherosclerotic disease and a case of

asymptomatic lymphoma. Wearable sensors are emerging as a transformative technology for precision health and medicine and heart rate monitoring led to the diagnosis of atrial fibrillation, sleep apnea and detection of Lyme disease in participants. Large population-based initiatives such as “myHeart counts” are evaluating the potential of wearable heart sensors to detect subclinical atrial fibrillation[7] and electrocardiographic monitoring is now available in consumer wearable devices[64]. Our findings also suggest a role for CGM in diabetes prevention by identifying unrecognized glucose dysregulation[6], and enabling individual to optimize diet based on personalized glycemic responses.

Our multi-omics analysis also provided important insights into ASCVD risk, highlighting the importance of systemic inflammation. Although our study was not powered for outcome analysis, all 5 participants with incident cardiovascular events had subclinical inflammation. Furthermore, correlation network analysis highlighted the role of monocytes, HGF, IL-2, MCP-3 and interferon-gamma cytokines including MIG and IP10 and other molecules in cardiovascular health. These analytes are involved in inflammation and are emerging in the context of ASCVD[40-42,44,65].

Untargeted longitudinal outlier analysis of the period leading up to the diagnosis of lymphoma illustrates the importance of longitudinal multi-omics analysis for biomarker and pathway discoveries. We identified potential critical biomarkers (*e.g.* MIG) and changes in the microbiome up to 1 year prior to diagnosis demonstrating the power of monitoring molecules longitudinally to detect deviations from the healthy baseline. Outlier biomarkers at time of diagnosis illustrated deregulated pathways related to inflammation, cell proliferation and cell migration that shed light on underlying dysregulated biological mechanisms associated with the disease. Further work will be needed to streamline the investigation of untargeted discoveries within precision medicine research. Given the need for early biomarkers for cancer detection, longitudinal multi-omics analyses represent an important tool for meeting this need. In addition to individual molecule monitoring, omics profiles provide the opportunity to detect outliers relative to a matched-healthy population. Clinical outlier analysis identified one participant with MGUS where early diagnosis with follow-up can increase survival time in individuals who progress to an associated malignancy[56]. While some omics outlier profiles could be clearly connected to an underlying health condition, the case of the participant with significant RNA-seq outliers illustrates the challenges of interpreting the clinical relevance of outlier analysis results with emerging measures. While precision medicine approaches have the potential for unnecessary anxiety and overtesting, we did not observe this in our population.

In the rapidly evolving field of precision medicine, this study should be assessed in the context of methodological considerations. Our cohort comprised highly educated volunteers, and therefore likely had a self-selection bias. Although this may affect the generalizability of our findings for behavioral changes, it is less likely to affect the underlying biological associations of multi-omics with glucose measures. A study strength is its ethnic diversity, which is greater than other longitudinal multi-omics studies[4,5]. We demonstrate the feasibility of a longitudinal precision health and medicine approach that builds on sound molecular and physiological phenotyping. We show that in-depth physiological and multi-omics characterizations is likely to further refine risk stratification. The intensive

longitudinal study design demonstrates how a small longitudinal cohort can yield important health and discovery findings. In the future, it will be possible to design personalized testing programs based on individual disease risk and longitudinal marker trajectories as well as evaluate the cost-value of these approaches for individuals and health care systems.

Data Availability

Raw omics data (transcriptome, immunome, proteome, metabolome, microbiome) included in this study are hosted on the NIH Human Microbiome 2 project site (https://portal.hmpdacc.org/) under the T2D project along with clinical laboratory data through 2016. Data from participants who have not consented to make their data public are available on dbGAP (accession phs001719.v1.p1). Additional data unique to this manuscript has been provided in supplemental data files.

Online Methods

Participant Consent and Accrual

Participants were recruited from the Stanford University surrounding community with the goal of enriching the cohort with individuals at risk for Type 2 diabetes and thus included individuals who expressed interest in other studies related to diabetes. Participants were enrolled as part of Stanford's iPOP (Integrated Personal Omics Profiling) research study (IRB 23602), which entails longitudinal multi-omics profiling of a cohort of adult volunteers enriched for pre-diabetes. There was no payment required to participate in the study and participants were not paid for their time. This study is part of the NIH integrated Human Microbiome Project (iHMP).

Design, Setting and Participants

The iPOP study is a longitudinal prospective cohort study[68] containing 109 individuals (Extended Data Figure S1a). Inclusion criteria were ages 25 to 75, body mass index (BMI) between 25 and 40 kg/m2 and 2-hour oral glucose tolerance test in the normal or prediabetic range (< 200 mg/dl). Exclusions included active eating disorder, hypertriglyceridemia > 400 mg/dL, uncontrolled hypertension, heavy alcohol use, pregnancy/lactation, prior bariatric surgery, and active psychiatric disease. After meeting initial recruitment goals, we expanded our inclusion criteria to include people with diabetes and people with normal BMI into the study. Participant demographics are summarized in Table 1 with detailed data provided in Tables D1, D2 and D3. Of note our cohort is slightly different than the main iHMP paper (Zhou et al., submitted). We excluded one participant who had no clinical history or follow-up information available and included 4 participants with clinical discoveries who entered the study after 2016 and thus had no omics data available.

The cohort was recruited over a number of years with the first participant starting in 2010. The study design has been described in detail previously[68]. Briefly, participants were asked to donate samples (*i.e.* fasted blood and stool) quarterly when healthy and more frequently when sick (viral infection), after immunization and various other events such as after taking antibiotics and going through colonoscopy. Samples collected through December 2016 were used for multi-omics analysis and corresponds to a median participation duration of 2.8

years. Standard and enhanced clinical lab data and participant surveys were available through June 2018. Most analysis were performed using healthy time points only. It is detailed in the text if all time points were used.

Measurements

All blood samples were collected after an overnight fast and were used to perform standard and enhanced clinical tests as well as emerging assays (Fig. 1). Standard tests included: FPG, HbA1C, fasted insulin, basic lipid panel, complete metabolic panel, CBC with differential and others (Table S1). In addition, participants were asked to complete various surveys in relation to demographics and current and past medical history, medications, smoking history, and family history, anthropometry, diet and physical activity as well as stress. Enhanced tests included: OGTT, SSPG, beta-cell function assessment, hsCRP, IgM, cardiovascular imaging (echocardiography, vascular ultrasound), cardiopulmonary exercise, CVD markers and wearable devices (physiology and activity monitor, continuous glucose monitoring (CGM)). In addition, multi-level molecular profiling were performed (emerging tests) including genome, gene expression (transcriptome), immune proteins (immunome), proteins (proteome), small molecules (metabolome), and gut microbes (microbiome). Clinical laboratory measures, immune proteins and cardiovascular biomarkers are detailed in Table S1. Participant surveys included the International Physical Activity Questionnaire, Stress and Adversity Inventory, and Perceived Stress Scale-10[69-71].

Modified Insulin Suppression Test

Sixty-nine participants underwent the modified insulin suppression test[72] to determine steady-state plasma glucose (SSPG) levels. The test was performed after an overnight fast and consists of 180-minute infusion of octreotide (0.27μg/m2/min), insulin (0.25 μg/m2/min), and glucose (240 μg/m2/min) with blood draws at minutes 150, 160, 170, and 180. The oximetric method was used to determine blood glucose and steady-state plasma glucose (SSPG) was determined by taking the mean of the four measurements. Reasons for not participating in this test included medical contraindications (n = 9), refusal (n = 5) and dropped out of study (n = 11) and not yet performed (n = 15).

Multi-omics Measures

Detailed methods regarding sample preparation, data acquisition and data preprocessing are available in the main NIH integrated Human Microbiome Project study by Zhou et al (submitted). We briefly summarize these methods here.

Genomics—Whole Exome Sequencing (n = 88) was performed by an accredited facility and variant calling was performed using an in-house pipeline (HugeSeq)[73]. Exomes were assessed for pathogenic variants according to the American College of Medical Genetics Guidelines[11,74]. The Online Mendelian Inheritance in Man (OMIM) database was used. Further details on processing and variant calling are provided in Rego et al.[11]

Peripheral Blood Mononuclear Cell (PBMC) RNA Sequencing—RNA sequencing from bulk PBMCs was performed using the TruSeq Stranded total RNA LT/HT Sample Prep Kit (Illumina) and sequenced on Illumina HiSeq 2000 instrument. The ‘TopHat’ package[75]

(v. 2.0.11) in R (v. 3.4) was used to align the reads to personal genomes, followed by 'HTseq' (v. 0.6.1) and 'DESEQ2'[76] (v. 3.5) for transcript assembly and RNA expression quantification.

Plasma SWATH-Mass Spectroscopy Proteomics—A NanoLC 425 System (SCIEX) was used to separate tryptic peptides of plasma samples. MS analyses were performed with randomized samples using SWATH Acquisition on a TripleTOF 6600 System equipped with a DuoSpray Source and 25 μm I.D. electrode (SCIEX). A final data matrix was produced with 1% FDR at peptide level and 10% FDR at protein level. Protein abundances were computed as the sum of the three most abundant peptides (top3 method). To address batch effects, subtraction of the principal components showing a major batch bias was performed using Perseus (v. 1.4.2.40).

Immune Protein Measurements—The 62 plex-Luminex antibody-conjugated bead capture assay (Affymetrix) was used to characterize blood levels of immune proteins. The assay was performed by the Stanford Human Immune Monitoring Center. The protocol is available at:http://iti.stanford.edu/content/dam/sm/iti/documents/himc/protocols/LuminexMultiplexAnalysisprotocol030213.doc (accessed May 1, 2018).

Plasma Liquid Chromatography-Mass Spectrometry (LC-MS) Metabolomics—Untargeted plasma metabolomics was performed using a broad spectrum LC-MS platform[77]. This analytical platform has been optimized to maximize metabolome coverage and involves complementary reverse-phase liquid chromatography (RPLC) and hydrophilic interaction liquid chromatography (HILIC) separations. Data were acquired on a Q Exactive plus mass spectrometer (Thermo Scientific) for HILIC and a Thermo Q Exactive mass spectrometer (Thermo Scientific) for RPLC. Both instruments were equipped with a HESI-II probe and operated in full MS scan mode. MS/MS data were acquired at various collision energies on pooled samples. LC-MS data were processed using Progenesis QI (Nonlinear Dynamics) and metabolic features were annotated by matching retention time and fragmentation spectra to authentic standards or to public repositories. Some metabolites elute in multiple peaks and are indicated with a number in parenthesis following the metabolite name ordered by elution time.

Plasma Lipidomics—Lipids were extracted and analyzed as previously described[78]. Briefly, we used a mixture of MTBE, methanol and water to extract lipids from 40 μl of plasma following biphasic separation. Lipids were then analyzed with the Lipidyzer platform consisting in a DMS device (SelexION Technology, SCIEX) and a QTRAP 5500 (SCIEX). Lipids were quantified using a mixture of 58 labeled internal standards provided with the platform. Lipodomics data is provided in Table D4.

16S Microbiome Sequencing—DNA was extracted from stool in line with the Human Microbiome Project's (HMP) Core Sampling Protocol A (hmpdacc.org). Targeted rRNA gene amplification of the V1 through V3 hypervariable regions of the 16S rRNA gene was performed using primers 27F and 534R (27F:5'-AGAGTTTGATCCTGGCTCAG-3' and 534R: 5'- ATTACCGCGGCTGCTGG-3'), and subsequently sequenced using 2×300 bp paired-end sequencing (Illumina MiSeq). Illumina's software handles initial processing of

all the raw sequencing data. A standard of one mismatch in primer and zero mismatch in barcode was applied to assign read pairs to the appropriate sample within a pool of samples. Barcodes and primers were removed prior to analysis. Amplicon sequences were clustered and Operational Taxonomic Units (OTU) picked by Usearch against GreenGenes database (May 2013 version) and final taxonomic assignment were performed using RDP-classifier.

ASCVD Circulating Markers

Millipore immunoassays human cardiovascular disease panels 1 to 4 (HCVD1MAG-67K, HCVD2MAG-67K, HCVD3MAG-67K, HCVD4MAG-67K) were used to characterize blood ASCVD circulating markers. The assays were performed by the Stanford Human Immune Monitoring Center.

Wearable Physiology and Activity Monitoring

Participants wore a Basis watch during the first part of the study and a Fitbit Charge 2 during the latter part of the study. We developed a special algorithm, “Change of Heart” to detect abnormalities in heart rate relative to a person’s baseline which was shown to provide an early warning signal of clinical abnormalities and disease which is described in detail in Li et al[2].

Continuous Glucose Monitoring

Continuous glucose monitoring (CGM) was performed with the Dexcom G4 CGM system. Participants wore the monitors for 2–4 weeks with interstitial glucose concentrations recorded every 5 minutes. They were also given glucose meters (AccCheck Nano SmartView) to measure finger prick blood glucose concentrations twice a day for the purpose of calibration.

Echocardiography

Baseline rest echocardiography was performed using commercially available echo systems (iE33; Philips Medical Imaging, Eindhoven, the Netherlands). Post-stress images were acquired immediately post-exercise, as per international consensus. Digitized echocardiographic studies were analyzed by the Stanford Cardiovascular Institute Biomarker and Phenotypic Core Laboratory on Xcelera workstations in accordance with published guidelines of the American Society of Echocardiography[79]. Regarding specific echocardiographic variables, left ventricular ejection fraction (LVEF) was calculated by manual contouring of apical imaging[80]. Left ventricular global longitudinal strain (LV GLS) was calculated from triplane apical imaging on manual tracings of the mid wall with the formula for LaGrangian Strain % = 100 x (L1 - L0)/L0), as previously described[81]. With tissue Doppler imaging, we used peak myocardial early diastolic velocity at the lateral mitral annulus and the assessment of trans mitral to tissue Doppler imaging early diastolic velocity ratio (E/e’)[82,83].

Vascular Ultrasound

Screening for subclinical atherosclerosis was performed using vascular ultrasound of the carotid and femoral artery using a 9.0 MHz Philips linear array probe and iE33 xMATRIX

echocardiography System manufactured by Philips (Andover, MA, USA). Vascular stiffness was assessed using central pulse wave velocity (PWV).

Cardiopulmonary Exercise Testing

Symptom-limited cardiopulmonary exercise (CPX) ventilatory expired gas analysis was completed with an individualized RAMP treadmill protocol[84]. Participants were encouraged to exercise to maximal exercise capacity. In addition, we monitored the respiratory exchange ratio (RER) during exercise and considered an RER ratio < 1.05 as representing sub-optimal or limitations associated with fatigue. Ventilatory efficiency (VE), oxygen consumption (VO_2), volume of carbon dioxide production (VCO_2) and other CPX variables were acquired breath by breath and averaged over 10 second intervals using CareFusion Oxygen Pro (San Diego, California) or CosMEd Quark (Rome, Italy) metabolic system. VE and VCO_2 responses throughout exercise were used to calculate the VE/VCO_2 slope via least squares linear regression ($y = mx + b$, m = slope)[85]. Percent predicted maximal oxygen consumption was derived using the Fitness Registry and the Importance of Exercise: a National Database (FRIEND) registry equation, derived from a large cohort of healthy US individuals who completed cardiopulmonary exercise testing[86].

iPOP Participant Surveys

Participants completed a survey on how the study had impacted their eating and exercise habits, what they learned about their health during the study, whether they discussed findings with their doctor, any follow-up testing, and other people they shared data with. This survey was initially administered anonymously but we then switched to surveys identified by participant ID. The quantitative results reported in Fig. 6 are from all participants who filled out an identifying survey (using last filled out survey where there were more than one). We used participant comments from anonymous and identified surveys in Table S27. At each quarterly visit, participants were asked about changes to health and medication. Participants were also asked by the study dietician how iPOP participation and CGM monitoring impacted their health behaviors (Table S28).

Calculation of Insulin Secretion Rate and Disposition Index

We used the ISEC program[87] to calculate the insulin secretion rate (ISR) from deconvolution of c-peptide measurements from plasma sampled at various time points during the OGTT (at minutes 0, 30 and 120). The deconvolution method uses population-based kinetic parameters[14] for c-peptide clearance to estimate insulin secretion rates at other timepoints. ISR was reported in pmol/kg/min at every 15-minute time interval between 0 and 120 minutes. The disposition index (DI) was calculated as the ISR at 30 minutes (ISR30) times the Matsuda index, which was calculated as in Cersosimo et al[13]. DI was reported as (pmol/kg/min)/(mg/dL*μU/mL).

Cluster Analysis and Association of Disposition Index with Multi-omics Measures

Insulin secretion rates were row standardized across the 9 timepoints from an OGTT sample and then clustered via the k-means clustering algorithm in R (v. 3.5) (function ‘kmeans’), with k = 4. Simple linear models were used to associate the disposition index with each

multi-omics analyte. Values for multi-omics analytes were from the time point closest to the OGTT date. Adjustment of p-values for multiple testing was performed using the Benjamini-Hochberg method, with an adjusted p-value of < 0.10 used to identify analytes significantly associated with the disposition index.

ASCVD and Adjusted ASCVD Risk Score Calculation

The ASCVD Pooled Cohort Risk Equations were implemented according to the instructions in the 2013 ACC/AHA Guideline on the Assessment of Cardiovascular Risk[33], using SAS 9.4 statistical software. The baseline time point was used for all participants except those that turned 40 during the study. In these cases, the first time point after age 40 was chosen. Participants under the age of 40 ($n = 7$) for the entire duration of the study were assigned the age of 40 for the purposes of ASCVD risk score calculation. To calculate the optimal risk for someone of a particular, age, sex and race, we used total cholesterol of 170, HDL of 50, and systolic blood pressure of 110 with no blood pressure medications, diabetes, or smoking. Adjusted ASCVD risk score was calculated by subtracting the optimal ASCVD risk score for a person of the same age, gender and race, from the participant's ASCVD risk score.

Association of Multi-omic Analytes and Adjusted ASCVD Risk Score

First, a median value was calculated for each analyte in each participant using healthy time points. A minimum of three healthy visits per participant was required. Spearman correlations were then calculated between adjusted ASCVD risk score and the median value of each multi-omics analyte. Associations were considered significant for analytes with q-value < 0.2. FDR correction was performed using the 'qvalue' package (v. 1.36.0) in R (v. 3.0.1).

Correlation Network Analysis

Spearman correlations among molecules significantly associated with disposition index and adjusted ASCVD risk score were calculated using the rcorr function in the 'Hmisc' package (v. 3.15–0) in R (v. 3.0.1) and p-values were corrected for multiple hypothesis using Bonferroni. Correlation networks were plotted using the R package 'igraph' (v. 0.7.1) and the layout used was Fruchterman-Reingold. Edges represent correlations with Bonferroni-corrected p-value < 0.05 and 0.10 for the disposition index and ASCVD risk score, respectively.

Linear Mixed Models (healthy-baseline and dynamic models)

SAS 9.4 Proc Mixed was used to perform linear mixed model analysis using the full maximum likelihood method of estimation and the between-within method for estimating degrees of freedom. We used a random intercept model with an unstructured covariance matrix for all analytes. Since linear time explained only a small amount of within person variation in FPG (1.2%) and HbA1C (5.0%) at healthy timepoints, we did not include time in our models. The outcome measures (FPG, HbA1C and hsCRP) were log-transformed in all models and the analytes were standardized to a mean of zero and standard deviation of one. All models were controlled for sex and age at consent. The healthy-baseline models used data from healthy quarterly visits. The dynamic analysis used the ratio to the first

available time point for each outcome measure and analytes and used all time points in the study. P-values were corrected for multiple hypothesis testing using the Benjamini-Hochberg procedure. Significant analytes have BH FDR < 0.2.

Data Reporting

In reporting results we considered consistency between models and results, validation through literature review of emerging molecules and relevance to disease state or risk condition. We also considered whether differing results varied because of sensitivity and variability of measures, the difference between evaluating absolute baseline values versus relative change, and the potential for biological saturation.

Multi-omics Outlier Analysis

Z-scores (mean of zero and standard deviation of one) were calculated after log2-transformation for all measures in all participants and outliers were defined as absolute Z-score > 95th percentile. Associated P-values were calculated assuming a normal distribution. P-values were corrected for multiple hypothesis using the Benjamini-Hochberg procedure.

Stroke Genes Outlier Analysis

Z-scores were calculated as described above for 14 of 32 genes recently identified as being associated with stroke and stroke types[39]. The 14 genes that we detected in our RNA-seq dataset were as follows: CASZ1, CDK6, FURIN, ICA1L, LDLR, LRCH1, PRPF8, SH2B3, SH3PXD2A, SLC22A7, SLC44A2, SMARCA4, ZCCHC14, ZFHX3. A composite Z-score was calculated by summing the individual gene Z-scores.

Pathway Enrichment Analysis

The web tool IMPaLA version 11 (build April 2018) (Integrated Molecular Pathway-Level Analysis) (http://impala.molgen.mpg.de) was used for the joint pathway analysis of proteins (from SWATH-MS) and metabolites (from LC-MS) abundances. Uniprot and HMDB accession numbers were used for proteins and metabolites, respectively. Pathway significance for proteins and metabolites separately was calculated using a hypergeometric test; the whole space of proteins and metabolites described in the pathways were used as a background. Joint p-values combining protein and metabolite pathways are calculated using Fisher's method. Multiple comparisons are controlled for using the Benjamini-Hochberg procedure[88].

Exercise Sub-study Analysis

ASCVD risk scores were calculated using cholesterol labs closest to the exercise study date using the same method as that used for the baseline ASCVD risk scores. Correlation analysis was done with 'corrplot' package in R (v. 3.3.2). The network was plotted using Cytoscape 3.4.0[89], where edges represent correlations with statistically significant Spearman's values (FDR < 0.2). False discovery rate correction was performed using the 'qvalue' package (v. 1.36.0) in R. The distance between nodes represents the strength of the pull between a node and its connected neighbors. The larger the value, the closer the distance between the two

nodes. The system was iterated until dynamic equilibrium using the prefuse force directed layout[90].

Microbiome Diversity: Univariate Models

Shannon diversity was calculated with SAS 9.4 using a code adapted from Montagna[91]. SAS 9.4 Proc Mixed using restricted maximum likelihood estimation the between-within degrees of freedom method was used to model the association of HbA1c, FPG and SSPG and Shannon diversity H' index. Preliminary analyses were done in proc gam and suggested an 'inverse u' distribution for all 3 measures in relationship to the Shannon diversity index. HbA1C and FPG were modeled using a repeated measures model with spatial power covariance structure. Shannon was entered into the model as a quadratic predictor of HbA1C and FPG. SSPG was modeled slightly differently because SSPG was only measured once in participants thus models with the predictor SSPG included Shannon diversity in the random statement. In addition, Shannon diversity as a quadratic term did not improve model fit and was not significant in any SSPG models so we present only the models with Shannon as a linear predictor (Table 6).

Microbiome Diversity: Multivariate Model

For our multivariate model (SAS 9.4 Proc Mixed), the full maximum likelihood method of estimation was used to enable comparison between models. The degree of freedom method was the between-within method. We used an unstructured covariance matrix for the models presented. In addition to the models presented in Table S7, we also evaluated the effect of adding of baseline BMI, consent age, or metformin use to the model. None of these covariates added significantly to the model and thus were left out of subsequent models. In addition, we evaluated whether use of the Firmicutes/Bacteroidetes ratio in place of the phylum Bacteroidetes proportion would improve the model. However the ratio accounted for substantially less within person variation in Shannon diversity (10.4%) thus we kept the proportion of the phylum Bacteroidetes in the final model.

Modeling Individual Shannon Diversity Trajectories

We modeled the change in Shannon diversity over time for individual participants using a general additive model (SAS proc gam) which separates the linear and non-linear components of the trajectory. The F test of the model using time as a predictor of Shannon diversity was compared to the null model and was calculated according to SAS usage note 32927:http://support.sas.com/kb/32/927.html (accessed March 2018).

SSPG and OGTT prediction models

Reprocessing of microbiome data—For the prediction models, the microbiome 16S reads were reprocessed using QIIME 2[92] (https://qiime2.org) and the DADA2[93] denoising plugin. The resulting read depth was 18,885 ± 11,852 (mean ± SD) following paired-end joining, removal of chimeric reads, and removal of samples with <7000 read depth. Taxonomic assignment was carried out using a naïve Bayes classifier trained on the above primers with the 99% 13_8 Greengenes OTU data set as reference sequences[94]. DADA2 facilitates cross-study comparison by providing DNA sequences of features thus making it

more appropriate for prediction models which will eventually need further external validation[95].

Feature selection—Features from multi-omics (clinical labs, transcriptome, immunome, proteome, metabolome, lipidome and microbiome) were standardized to zero mean with unit variance. Clinical laboratory (including SSPG), immunome and metabolomics data was log transformed prior to standardization. The variance stabilizing transformation had been used for RNA-seq data. The sample IDs used for each SSPG and OGTT model are provided in Data Tables D5-D24. We then used the 'MXM' R package[26] (v. 0.9.7) with the Max-Min Parents and Child algorithm (MMPC)[25] option to identify features that are parents or children of SSPG in a Bayesian network constructed from all the available data. The features selected by the algorithm are hypothesized to be direct causes or effects of SSPG in the data, as each feature selected are SSPG dependent when conditioned on every possible subset of the other features. These features provide novel information about SSPG, and thus are most useful for prediction. There were 41 participants with SSPG values and all multi-omics data. Feature selection was performed using leave-one-out cross validation, where 41 training sets were constructed and each training set excludes the data from a different patient. We ran the MMPC algorithm on each training set. Features that were identified by the MMPC algorithm in ≥ 20% of training sets were used as features in the model. For the OGTT predictive model, there was no lipidomics data available.

Ridge Regression—Ridge Regression was performed using R (v. 3.4.1). For each -ome, we use the sample at the closest time point that is equal or prior to the time point of the patient's SSPG/OGTT measurement. We performed leave one out cross validation to maximize available training data. For each training set, we optimize the hyperparameter by performing a grid search and selecting the model that minimizes test error. The predicted SSPG/OGTT value is the value from the cross validation iteration in which that SSPG/OGTT data point and its associated features are excluded from the training set. We use these predicted values to calculate mean square error and R^2 values. The value of the hyperparameter used was the average of the hyperparameters which minimized test error during cross validation.

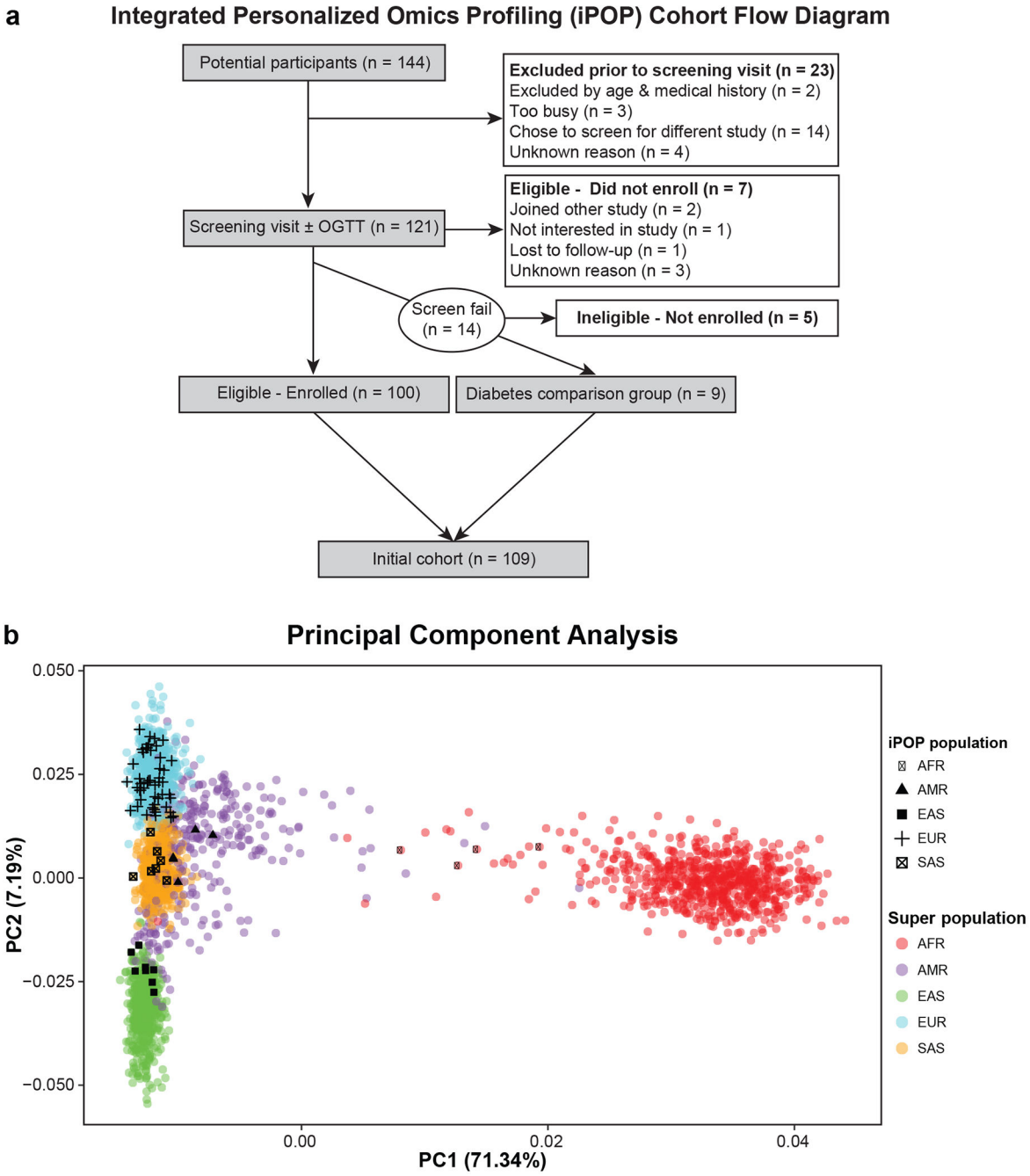
Ethnicity PCA Plot

Ethnicity information for 72 individuals in the study was broadly classified into the five 1000 Genomes Project (1000GP) Consortium super-population definitions, which are namely African (AFR), East Asian (EAS), European (EUR), South Asian (SAS) and admixed American (AMR). Individuals who self-identify as Indians from South Asia were categorized as SAS ($n = 7$), Hispanics and Latinos as AMR ($n = 3$), East Asians as EAS ($n = 8$), Caucasians as EUR ($n = 50$) and African Americans ($n = 4$) as AFR. The ethnicity information from the 2,504 samples, definitions of the populations and super-populations, and genetic information of the 1000GP were obtained from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/ (downloaded in April 2017).

The following filters were first implemented for each individual genome for the study: (a) we removed indels, leaving only the SNVs, (b) we removed SNVs without the "PASS" tag,

(c) we kept SNVs with a minimum read depth of 1, and (d) we removed SNVs with missing genotypes. We then intersected the genetic loci from 72 individuals and the samples from the 1000GP, to obtain 6,653 SNVs common to both datasets. In order to reduce the chance of linkage disequilibrium and dependency between SNVs due to close proximity, we further thin the SNV set by taking every third SNV. Finally, we have a combined set of 2,576 samples and 2,318 SNVs that we use for PCA. We used the smartpca tool in the PLINK2 suite to generate the PCA[96].

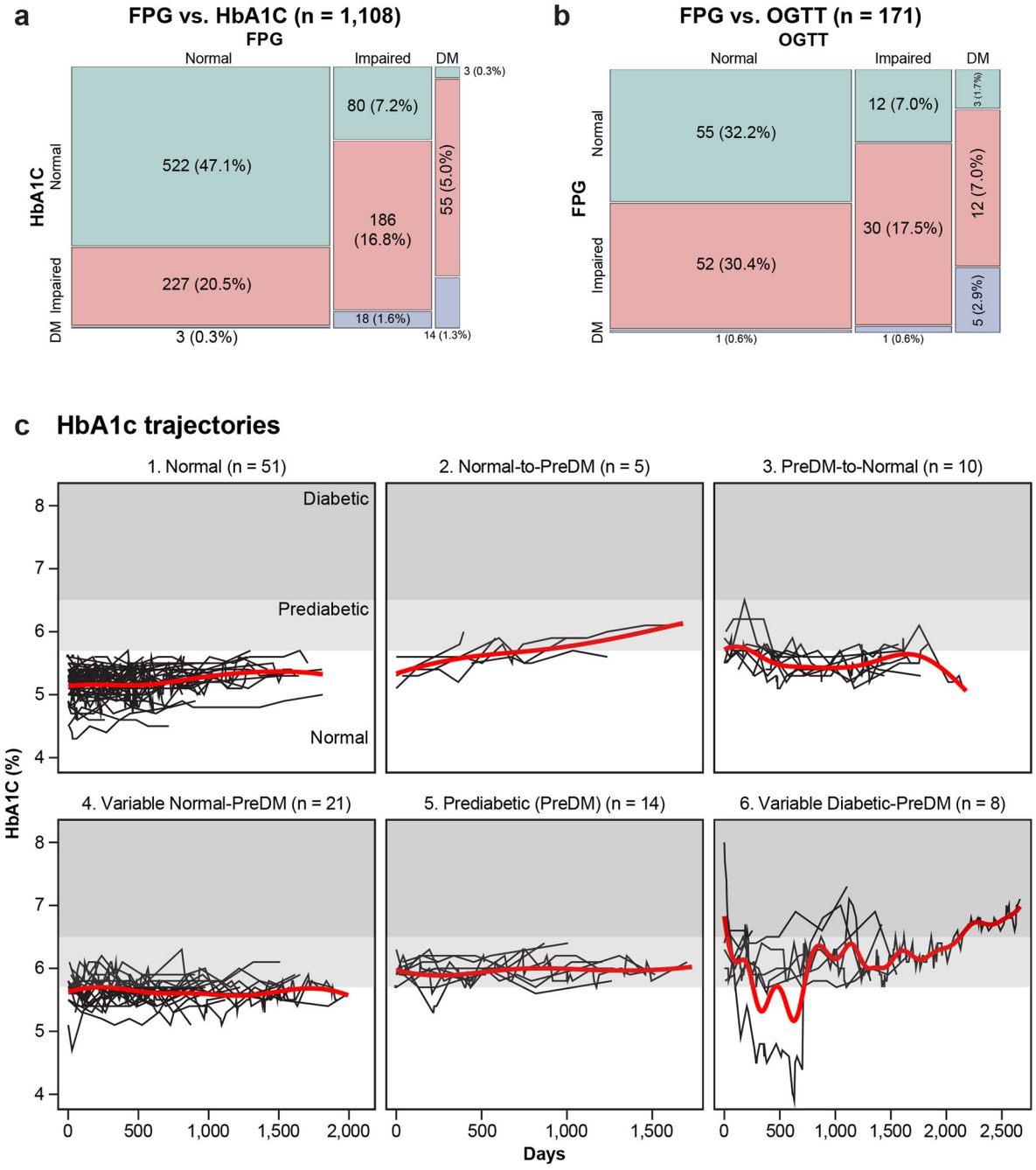
Extended Data



Extended Data Fig 1. Integrated personalized omics profiling cohort flow chart and genetic ancestry.

(**a**) The flow chart demonstrates recruitment and enrollment of the iPOP cohort. (**b**) Principal components analysis (PCA) plot showing the ancestries of 72 participants. The reference includes 2,504 samples from the 1000 Genomes Project[10]. Each filled circle is a 1000GP sample, colored by the super-population of ancestral origin, namely African (AFR; red), admixed American (AMR; purple), East Asian (EAS; green), European (EUR; cyan) and South Asian (SAS; orange). Each black symbol is an individual from the study, which we categorized by self-reported ethnicity consistent with the 1000GP super-population

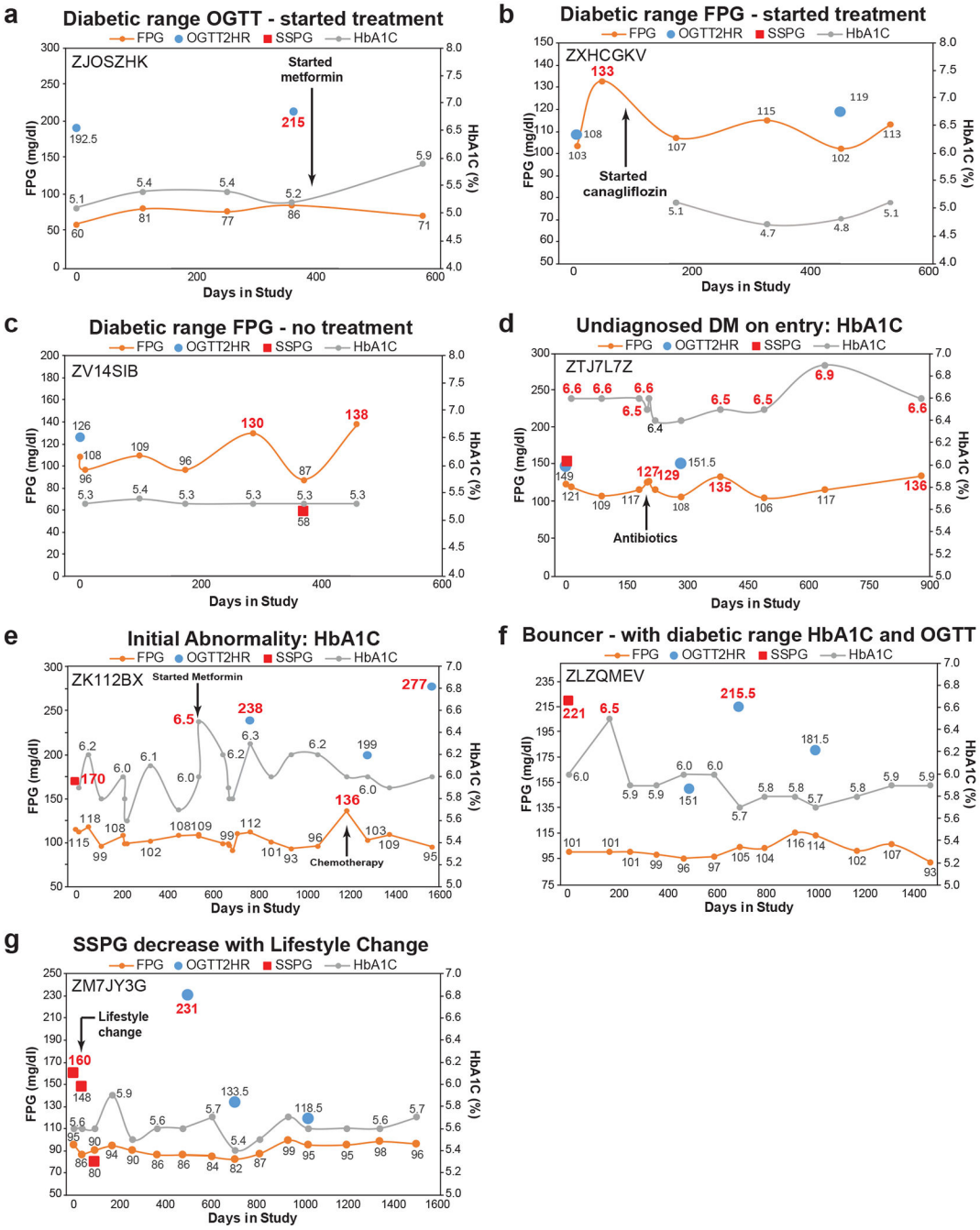
definitions, namely AFR (black filled circle), AMR (black filled triangle), EAS (black filled square), EUR (black plus sign) and SAS (a checked box). We see that the individuals in our study have self-reported ancestries generally clustering within the super-population reference panel from the 1000GP.



Extended Data Fig 2. Comparison of diabetic metrics in categorizing individuals when performed at the same time and HbA1C trajectories.

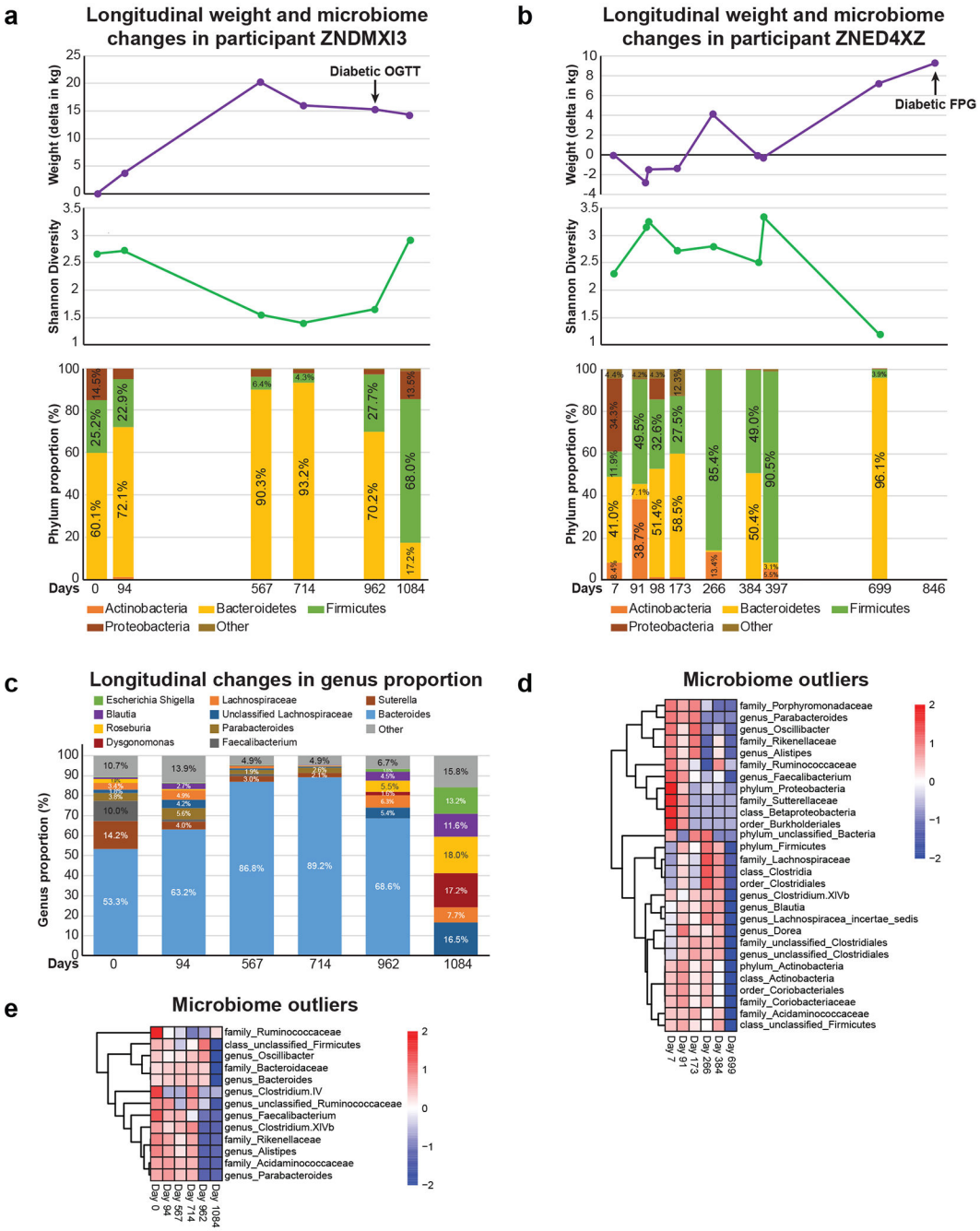
(**a**) Overlap of Fasting Plasma Glucose (FPG) and Hemoglobin A1C (HbA1C) categories when simultaneously measured. FPG impaired: 100 mg/dL ≤ FPG < 126 mg/dL; diabetic range: FPG ≥ 126 mg/dL; HbA1C impaired: 5.7% ≤ HbA1C < 6.5%; diabetic range: HbA1C ≥ 6.5%. (**b**) Overlap of FPG and 2-Hour Oral Glucose Tolerance Test (OGTT) when simultaneously measured. FPG ranges as above. OGTT impaired: 140 mg/dL ≤ OGTT < 200 mg/dL; diabetic range ≥ 200 mg/dL. (**c**) Longitudinal patterns of changes in Hemoglobin A1C (HbA1C) over time. Six different patterns could be characterized

including: 1- participants who remained in the normal range the entire study (Group 1, n = 51), 2- participants who progressed from normal to prediabetic (Group 2, n = 5), 3- participants who went from prediabetic to normal (Group 3, n = 10), 4- participants whose HbA1C went back and forth from normal to prediabetic (Group 4, n = 21), 5- participants whose HbA1C labs were predominantly in the prediabetic range (Group 5, n = 14), and 6- participants whose HbA1C crossed into the diabetic range (Group 6, n = 8). The red lines represent the overall penalized b-spline of participants' data in each category.



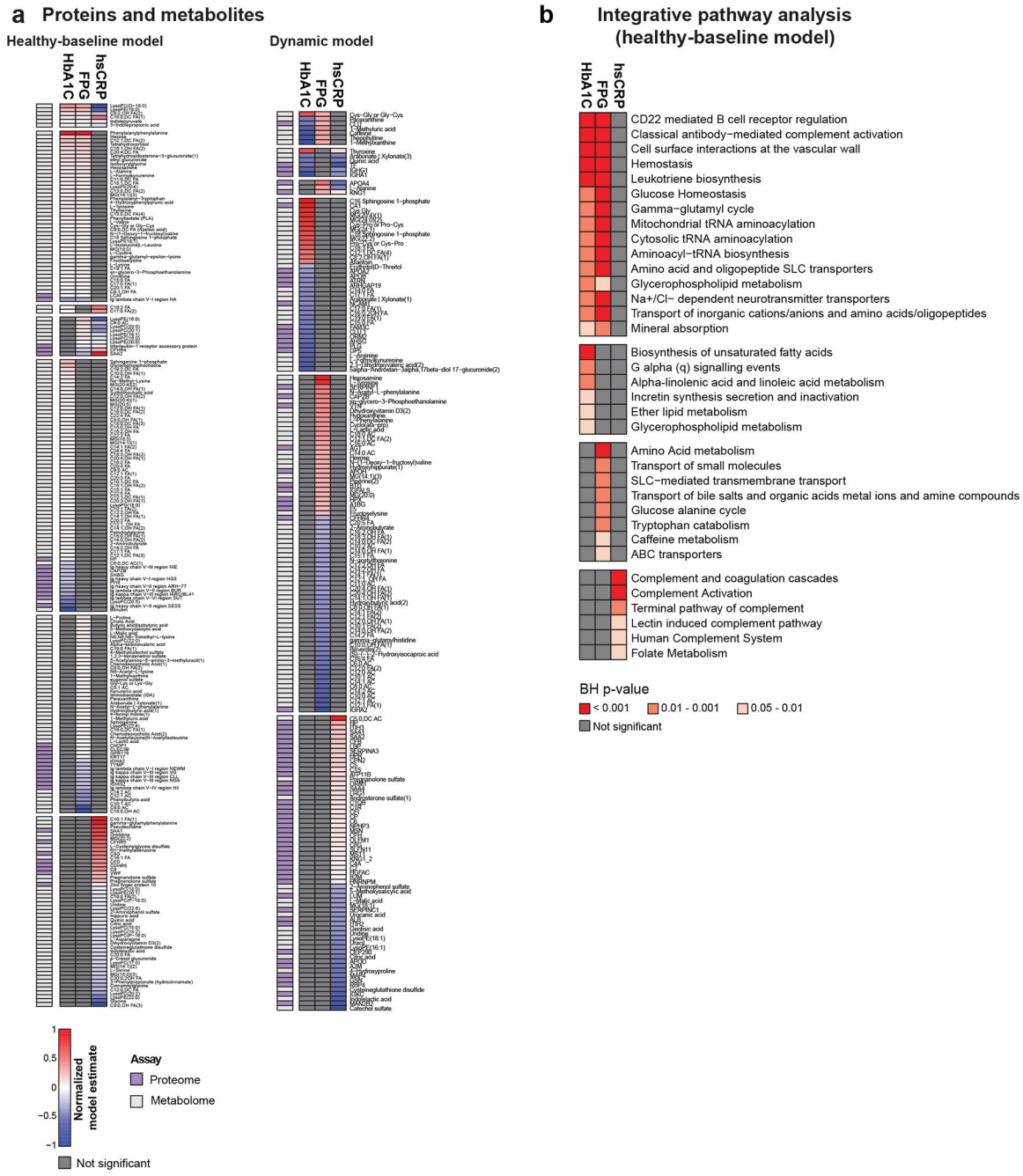
Extended Data Fig 3. Additional individual longitudinal trajectories for diabetic measures.

Diabetic-range metrics are indicated in red. (**a**) Diabetic range OGTT, (**b,c**) Diabetic range FPG, (**d**) undiagnosed DM at study entry (HbA1C), (**e**) Initial abnormality HbA1C. Note this person had two HbA1C measurements on the same day at two different laboratories and was started on medication based on the higher measurement, (**f**) Bouncer with diabetic range HbA1C and OGTT, and (**g**) SSPG decrease with lifestyle change.



Extended Data Fig 4. Longitudinal microbiome trajectories in diabetes.

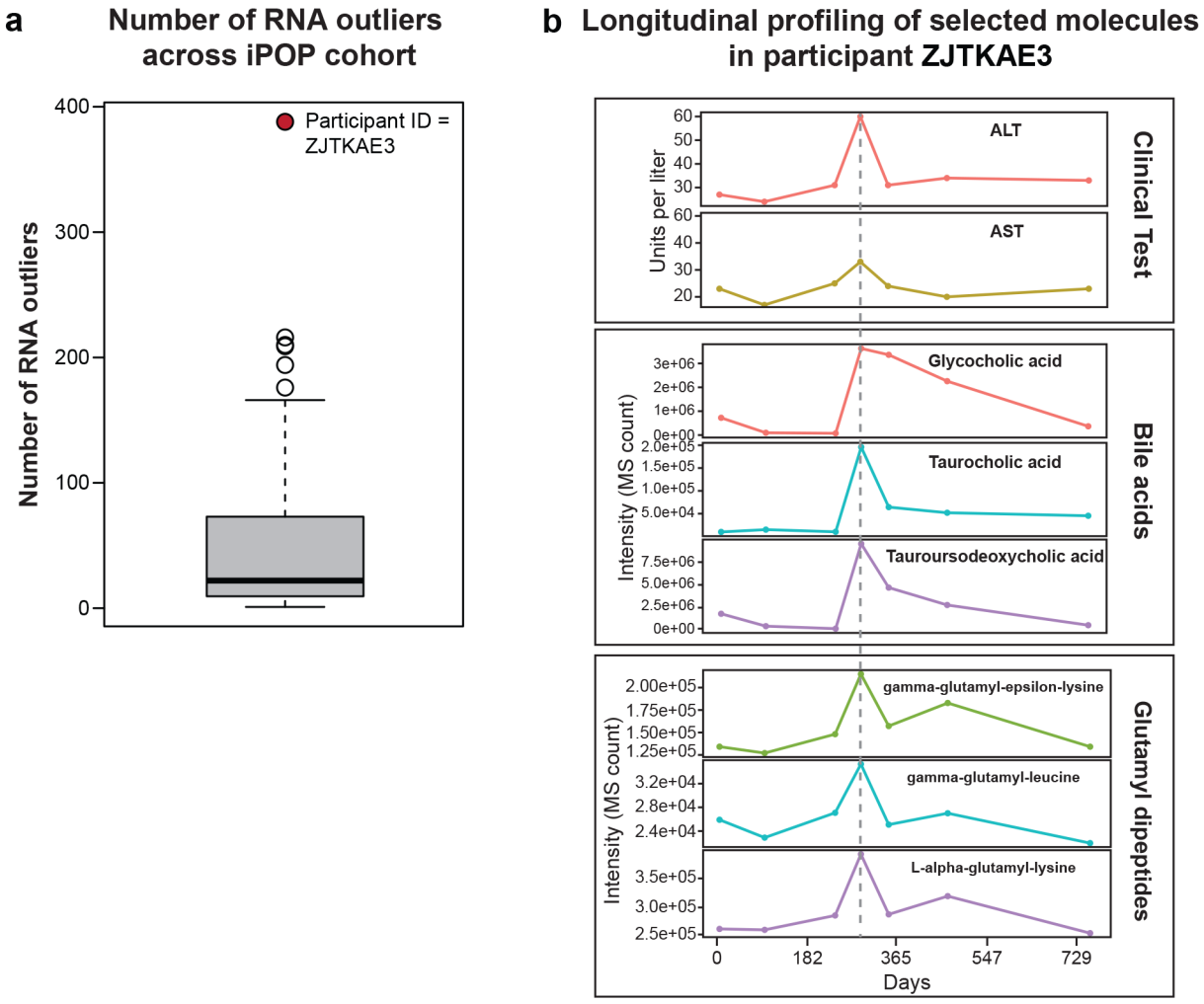
Longitudinal weight, gut microbial Shannon diversity and phylum proportion changes in participants (**a**) ZNDMXI3 and (**b**) ZNED4XZ. (**c**) Longitudinal changes in genus proportion (ZNDMXI3). Microbiome outliers (95th percentile) at the latest microbiome sample time point in participants (**d**) ZNDMXI3 and (**e**) ZNED4XZ. Microbial abundance is scaled by row with low (blue) and high (red) abundance.



Extended Data Fig 5. Multi-omics of glucose metabolism and inflammation.

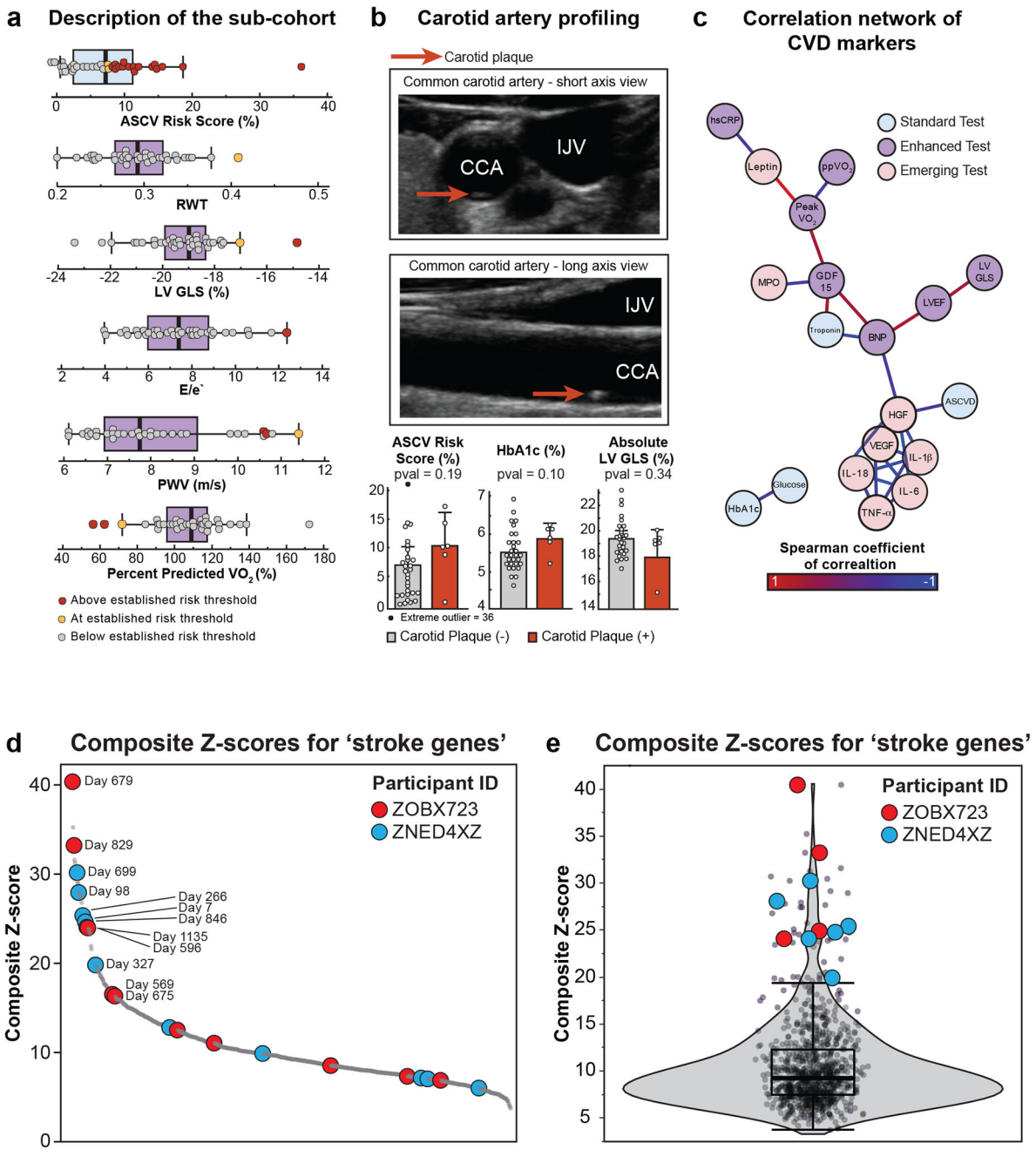
(**a**) Proteins and metabolites associated with HbA1C, FPG, and hsCRP using healthy-baseline and dynamic linear mixed models. Healthy-baseline models (HbA1C $n = 101$, samples 560; FPG $n = 101$, samples 563; hsCRP $n = 98$, samples 518) account for repeated measures at healthy time points. Dynamic models are similar models except that analytes are normalized across individuals to the first measurement and all time points in the study are used (HbA1C $n = 94$, samples = 836; FPG $n = 94$, samples = 843; hsCRP $n = 92$, samples 777). Individual analyte p-values were determined using a two-sided t-test. Multiple testing correction was performed and molecules were considered significant when BH FDR < 0.2.

Model estimates were normalized in each condition so the maximum value equal 1 and the minimal value equal −1. (**b**) Integrative pathway analysis using IMPaLa (http://impala.molgen.mpg.de) of proteins and metabolites associated with HbA1C (n = 101, samples 560), FPG (n = 101, samples 563), and hsCRP (n = 98, samples 518) as determined by the healthy-baseline models (BH FDR < 0.2 at molecule level which matched to known pathways. Significance of pathways for proteins and metabolites separately is determined by the hypergeometric test (one-sided) followed by Fisher's combined probability test (one-sided) to determine combined pathway significance (BH FDR < 0.05; n's of proteins and metabolites for each pathway are provided in Tables S9, S11, S13).



Extended Data Fig 6. Outlier Analysis of RNA-seq data.

(**a**) Number of outlier RNA molecules (95th percentile) in each participant. Outlier analysis was performed on Z-scores calculated on the median expression level of each gene at healthy visits in individuals with at least 3 healthy visits (n = 63). The box is defined as 25th and 75th quartile. The upper whisker extends to 1.5 times the interquartile range from the box and the lower whisker to the lowest data point. The horizontal bar in the box is the median value. (**b**) Selected clinical lab and metabolite trajectories (7 measurement time points) for participant ZJTKAE3 showing a concomitant increase of bile acids and glutamyl dipeptides with ALT (alanine aminotransferase) and AST (aspartate aminotransferase).



Extended Data Fig 7. Multidimensional cardiac risk assessment.

(**a**) Distribution of ASCVD risk scores (n = 35, 36 measurements) and cardiovascular imaging and physiology measures that have been established as cardiovascular risk markers. (Abbreviations: RWT-relative wall thickness, LV GLS-left ventricular global longitudinal strain, E/e' - ratio of mitral peak velocity of early filling (E) to early diastolic mitral annular velocity (e'), PWV-pulse wave velocity). Please note that thresholds for PWV are age-related. Box plots were derived to display quartiles (Q1, median, Q3) with the upper whisker being Q3 plus 1.5*(interquartile range) and the lower whisker extending to Q1 minus 1.5*(interquartile range) or the lowest data point. (**b**) Ultrasound of carotid plaque (6

participants of 36 had an ultrasound finding of carotid plaque) and relative distribution of ASCVD risk score, HbA1C and LV GLS in function of presence or absence of carotid plaque (Student's t-test (two-sided) was used to evaluate differences between groups; $n = 35$, 36 measurements). Error bars represent one standard deviation from the mean (upper edge of box). (**c**) Correlation network of selected metrics collected during cardiovascular assessment which associated (Spearman correlation (two-sided) with ASCVD risk score (q-value < 0.2); n = 35 participants with 36 measurements. (**d**) Composite Z-score of ZOBX723 (unstable angina with stent placement) and ZNED4XZ (mild stroke with full recovery and transition to diabetes). For ZOBX723, day 829 occurred 3 weeks post stent placement. Day 679 was a mid-infection time point. For ZNED4XZ, day 699 was the time point prior to the participant's transition to diabetes and day 846 was the first diabetic time point. The stroke occurred on day 307 for this individual. Gray dots represent Z-scores of other participants (n=101 with 859 samples). (**e**) Violin plot showing the same data as (d) ($n = 101$ with 859 samples). The box plot shows the 1st (lower edge of box), median (middle line) and 3rd (upper edge of box) quartiles. The upper whisker is the 3rd quartile + 1.5*(interquartile range) and the lower whisker is the lowest data point.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Our work was supported by grants from the National Institutes of Health (NIH) Human Microbiome Project (HMP) 1U54DE02378901 (G.M.W. and M.P.S.), an NIH grant no. R01 DK110186-03 (T.L.M.), a NIH National Center for Advancing Translational Science Clinical and Translational Science Award (no. UL1TR001085). This work used the Genome Sequencing Service Center by the Stanford Center for Genomics and Personalized Medicine Sequencing Center (supported by NIH grant no. S10OD020141), the Diabetes Genomics Analysis Core and the Clinical and Translational Core of the Stanford Diabetes Research Center (NIH grant no. P30DK116074). SMS-FR was supported by a Department of Veteran Affairs Office of Academic Affiliations Advanced Fellowship in Spinal Cord Injury Medicine and a NIH Career Development Award K08 ES028825. GMS was supported by NIH grant K08 MH103443. DH was supported by a Stanford School of Medicine Dean's Postdoctoral Fellowship and a Stanford Center for Computational, Evolutionary and Human Genomics Fellowship. MRS was supported by grants P300PA_161005 and P2GEP3_151825 from the Swiss National Science Foundation (SNSF). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH, the Department of Veteran Affairs, or the SNSF. We thank Songjie Chen and Brittany Lee for their work in metabolomics data production. Alessandra Breschi generously shared her code for the insulin secretion rate calculations. Finally, we thank the iPOP participants who generously gave their time and biological samples.

References

1. National Research Council (US) Committee on A Framework for Developing a New Taxonomy of Disease. Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease. (National Academies Press (US), 2012).
2. Li X et al. Digital Health: Tracking Physiomes and Activity Using Wearable Biosensors Reveals Useful Health-Related Information. PLoS Biol. 15, e2001402 (2017). [PubMed: 28081144]
3. Chen R et al. Personal omics profiling reveals dynamic molecular and medical phenotypes. Cell 148, 1293–1307 (2012). [PubMed: 22424236]
4. Price ND et al. A wellness study of 108 individuals using personal, dense, dynamic data clouds. Nat. Biotechnol. (2017). doi:10.1038/nbt.3870
5. Perkins BA et al. Precision medicine screening using whole-genome sequencing and advanced imaging to identify disease risk in adults. Proc. Natl. Acad. Sci. U. S. A. (2018). doi:10.1073/pnas.1706096114

6. Hall H et al. Glucotypes reveal new patterns of glucose dysregulation. PLoS Biol. 16, e2005143 (2018). [PubMed: 30040822]
7. McConnell MV et al. Feasibility of Obtaining Measures of Lifestyle From a Smartphone App: The MyHeart Counts Cardiovascular Health Study. JAMA Cardiol 2, 67–76 (2017). [PubMed: 27973671]
8. Dinneen S, Gerich J & Rizza R Carbohydrate metabolism in non-insulin-dependent diabetes mellitus. N. Engl. J. Med. 327, 707–713 (1992). [PubMed: 1495524]
9. Varghese RT et al. Mechanisms Underlying the Pathogenesis of Isolated Impaired Glucose Tolerance in Humans. J. Clin. Endocrinol. Metab. 101, 4816–4824 (2016). [PubMed: 27603902]
10. 1000 Genomes Project Consortium et al. A global reference for human genetic variation. Nature 526, 68–74 (2015). [PubMed: 26432245]
11. Rego S et al. High Frequency Actionable Pathogenic Exome Variants in an Average-Risk Cohort. Cold Spring Harb Mol Case Stud (2018). doi:10.1101/mcs.a003178
12. Pearson ER et al. Genetic cause of hyperglycaemia and response to treatment in diabetes. Lancet 362, 1275–1281 (2003). [PubMed: 14575972]
13. Cersosimo E, Solis-Herrera C, Trautmann ME, Malloy J & Triplitt CL Assessment of pancreatic β-cell function: review of methods and clinical applications. Curr. Diabetes Rev. 10, 2–42 (2014). [PubMed: 24524730]
14. Van Cauter E, Mestrez F, Sturis J & Polonsky KS Estimation of insulin secretion rates from C-peptide levels. Comparison of individual and standard kinetic parameters for C-peptide clearance. Diabetes 41, 368–377 (1992). [PubMed: 1551497]
15. Matsuda M & DeFronzo RA Insulin sensitivity indices obtained from oral glucose tolerance testing: comparison with the euglycemic insulin clamp. Diabetes Care 22, 1462–1470 (1999). [PubMed: 10480510]
16. Godsland IF, Jeffs JAR & Johnston DG Loss of beta cell function as fasting glucose increases in the non-diabetic range. Diabetologia 47, 1157–1166 (2004). [PubMed: 15249997]
17. Kanat M et al. The relationship between {beta}-cell function and glycated hemoglobin: results from the veterans administration genetic epidemiology study. Diabetes Care 34, 1006–1010 (2011). [PubMed: 21346184]
18. Iikuni N, Lam QLK, Lu L, Matarese G & La Cava A Leptin and Inflammation. Curr. Immunol. Rev. 4, 70–79 (2008). [PubMed: 20198122]
19. Hamilton JA GM-CSF in inflammation and autoimmunity. Trends Immunol. 23, 403–408 (2002). [PubMed: 12133803]
20. Reidy SP & Weber J Leptin: an essential regulator of lipid metabolism. Comp. Biochem. Physiol. A Mol. Integr. Physiol 125, 285–298 (2000). [PubMed: 10794958]
21. Guasch-Ferré M et al. Metabolomics in Prediabetes and Diabetes: A Systematic Review and Meta-analysis. Diabetes Care 39, 833–846 (2016). [PubMed: 27208380]
22. Twig G et al. White blood cells count and incidence of type 2 diabetes in young men. Diabetes Care 36, 276–282 (2013). [PubMed: 22961572]
23. Oliveira AG et al. The Role of Hepatocyte Growth Factor (HGF) in Insulin Resistance and Diabetes. Front. Endocrinol. 9, 503 (2018).
24. Mothe-Satney I et al. Adipocytes secrete leukotrienes: contribution to obesity-associated inflammation and insulin resistance in mice. Diabetes 61, 2311–2319 (2012). [PubMed: 22688342]
25. Tsamardinos I, Brown LE & Aliferis CF The max-min hill-climbing Bayesian network structure learning algorithm. Mach. Learn. 65, 31–78 (2006).
26. Lagani V, Athineou G, Farcomeni A, Tsagris M & Tsamardinos I Feature Selection with the R Package MXM: Discovering Statistically Equivalent Feature Subsets. Journal of Statistical Software, Articles 80, 1–25 (2017).
27. McLaughlin T et al. Use of metabolic markers to identify overweight individuals who are insulin resistant. Ann. Intern. Med. 139, 802–809 (2003). [PubMed: 14623617]
28. Nowak C et al. Protein Biomarkers for Insulin Resistance and Type 2 Diabetes Risk in Two Large Community Cohorts. Diabetes 65, 276–284 (2016). [PubMed: 26420861]

29. Apostolopoulou M et al. Specific Hepatic Sphingolipids Relate to Insulin Resistance, Oxidative Stress, and Inflammation in Nonalcoholic Steatohepatitis. Diabetes Care 41, 1235–1243 (2018). [PubMed: 29602794]
30. Gomez-Arango LF et al. Connections Between the Gut Microbiome and Metabolic Hormones in Early Pregnancy in Overweight and Obese Women. Diabetes 65, 2214–2223 (2016). [PubMed: 27217482]
31. Kwo PY, Cohen SM & Lim JK ACG Clinical Guideline: Evaluation of Abnormal Liver Chemistries. Am. J. Gastroenterol. 112, 18–35 (2017). [PubMed: 27995906]
32. Hu FB et al. Elevated risk of cardiovascular disease prior to clinical diagnosis of type 2 diabetes. Diabetes Care 25, 1129–1134 (2002). [PubMed: 12087009]
33. Goff DC Jr et al. 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. Circulation 129, S49–73 (2014). [PubMed: 24222018]
34. Kuznetsova T et al. Additive Prognostic Value of Left Ventricular Systolic Dysfunction in a Population-Based Cohort. Circ. Cardiovasc. Imaging 9, e004661 (2016). [PubMed: 27329778]
35. Wang TJ et al. Carotid intima-media thickness is associated with premature parental coronary heart disease: the Framingham Heart Study. Circulation 108, 572–576 (2003). [PubMed: 12874190]
36. Mitchell GF et al. Arterial stiffness and cardiovascular events: the Framingham Heart Study. Circulation 121, 505–511 (2010). [PubMed: 20083680]
37. Moneghetti KJ et al. Applying current normative data to prognosis in heart failure: The Fitness Registry and the Importance of Exercise National Database (FRIEND). Int. J. Cardiol 263, 75–79 (2018). [PubMed: 29525067]
38. Hall KT et al. Polymorphisms in catechol-O-methyltransferase modify treatment effects of aspirin on risk of cardiovascular disease. Arterioscler. Thromb. Vasc. Biol. 34, 2160–2167 (2014). [PubMed: 25035343]
39. Malik R et al. Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes. Nat. Genet. (2018). doi:10.1038/s41588-018-0058-3
40. Cross DS et al. Coronary risk assessment among intermediate risk patients using a clinical and biomarker based algorithm developed and validated in two population cohorts. Curr. Med. Res. Opin. 28, 1819–1830 (2012). [PubMed: 23092312]
41. Ma H, Calderon TM, Fallon JT & Berman JW Hepatocyte growth factor is a survival factor for endothelial cells and is expressed in human atherosclerotic plaques. Atherosclerosis 164, 79–87 (2002). [PubMed: 12119196]
42. Bell EJ et al. Hepatocyte Growth Factor Is Positively Associated With Risk of Stroke: The MESA (Multi-Ethnic Study of Atherosclerosis). Stroke 47, 2689–2694 (2016). [PubMed: 27729582]
43. Chen X & Devaraj S Monocytes from metabolic syndrome subjects exhibit a proinflammatory M1 phenotype. Metab. Syndr. Relat. Disord. 12, 362–366 (2014). [PubMed: 24847781]
44. Elkind MS et al. Interleukin-2 levels are associated with carotid artery intima-media thickness. Atherosclerosis 180, 181–187 (2005). [PubMed: 15823291]
45. Porez G, Prawitt J, Gross B & Staels B Bile acid receptors as targets for the treatment of dyslipidemia and cardiovascular disease. J. Lipid Res. 53, 1723–1737 (2012). [PubMed: 22550135]
46. Berry CE & Hare JM Xanthine oxidoreductase and cardiovascular disease: molecular mechanisms and pathophysiological implications. J. Physiol. 555, 589–606 (2004). [PubMed: 14694147]
47. Sane DC, Kontos JL & Greenberg CS Roles of transglutaminases in cardiac and vascular diseases. Front. Biosci. 12, 2530–2545 (2007). [PubMed: 17127261]
48. Wollert KC, Kempf T & Wallentin L Growth Differentiation Factor 15 as a Biomarker in Cardiovascular Disease. Clin. Chem. 63, 140–151 (2017). [PubMed: 28062617]
49. Klok MD, Jakobsdottir S & Drent ML The role of leptin and ghrelin in the regulation of food intake and body weight in humans: a review. Obes. Rev 8, 21–34 (2007). [PubMed: 17212793]
50. Charbonneau B et al. Pretreatment circulating serum cytokines associated with follicular and diffuse large B-cell lymphoma: a clinic-based case-control study. Cytokine 60, 882–889 (2012). [PubMed: 23010502]

51. Przewoznik M et al. Recruitment of natural killer cells in advanced stages of endogenously arising B-cell lymphoma: implications for therapeutic cell transfer. J. Immunother. 35, 217–222 (2012). [PubMed: 22421939]
52. Haabeth OAW et al. Inflammation driven by tumour-specific Th1 cells protects against B-cell cancer. Nat. Commun 2, 240 (2011). [PubMed: 21407206]
53. Ding Q et al. CXCL9: evidence and contradictions for its role in tumor progression. Cancer Med. 5, 3246–3259 (2016). [PubMed: 27726306]
54. Rolny C et al. HRG inhibits tumor growth and metastasis by inducing macrophage polarization and vessel normalization through downregulation of PlGF. Cancer Cell 19, 31–44 (2011). [PubMed: 21215706]
55. Johnson LDS, Goubran HA & Kotb RR Histidine rich glycoprotein and cancer: a multi-faceted relationship. Anticancer Res. 34, 593–603 (2014). [PubMed: 24510988]
56. Go RS, Gundrum JD & Neuner JM Determining the clinical significance of monoclonal gammopathy of undetermined significance: a SEER-Medicare population analysis. Clin. Lymphoma Myeloma Leuk. 15, 177–186.e4 (2015). [PubMed: 25445471]
57. Turesson I et al. Monoclonal gammopathy of undetermined significance and risk of lymphoid and myeloid malignancies: 728 cases followed up to 30 years in Sweden. Blood 123, 338–345 (2014). [PubMed: 24222331]
58. Ahlqvist E et al. Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. The Lancet Diabetes & Endocrinology (2018). doi: 10.1016/S2213-8587(18)30051-2
59. Cauwenberghs N et al. Relation of Insulin Resistance to Longitudinal Changes in Left Ventricular Structure and Function in a General Population. J. Am. Heart Assoc 7, (2018).
60. Piening BD et al. Integrative Personal Omics Profiles during Periods of Weight Gain and Loss. Cell Syst (2018). doi:10.1016/j.cels.2017.12.013
61. Whirl-Carrillo M et al. Pharmacogenomics Knowledge for Personalized Medicine. Clinical Pharmacology & Therapeutics 92, 414–417 (2012). [PubMed: 22992668]
62. Li J et al. Decoding the Genomics of Abdominal Aortic Aneurysm. Cell 174, 1361–1372.e10 (2018). [PubMed: 30193110]
63. Douglas PS et al. The Future of Cardiac Imaging: Report of a Think Tank Convened by the American College of Cardiology. JACC Cardiovasc. Imaging 9, 1211–1223 (2016). [PubMed: 27712724]
64. Buhr S Apple's Watch isn't the first with an EKG reader but it will matter to more consumers. TechCrunch (2018).
65. Omer W, Naveed AK, Khan OJ & Khan DA Role of Cytokine Gene Score in Risk Prediction of Premature Coronary Artery Disease. Genet. Test. Mol. Biomarkers 20, 685–691 (2016). [PubMed: 27689253]
66. Pathway analysis with transcriptomics and metabolomics data. Available at: http://impala.molgen.mpg.de/. (Accessed: 27th December 2018)
67. Szklarczyk D et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res. 43, D447–52 (2015). [PubMed: 25352553]
68. The Integrative Human Microbiome Project: Dynamic Analysis of Microbiome-Host Omics Profiles during Periods of Human Health and Disease. Cell Host Microbe 16, 276–289 (2014). [PubMed: 25211071]
69. Cohen S Perceived stress in a probability sample of the United States. (Sage Publications, Inc, 1988).
70. Slavich GM & Shields GS Assessing Lifetime Stress Exposure using the Stress and Adversity Inventory for Adults (Adult STRAIN): An Overview and Initial Validation. Psychosom. Med (2017). doi:10.1097/PSY.0000000000000534
71. Lee PH, Macfarlane DJ, Lam TH & Stewart SM Validity of the International Physical Activity Questionnaire Short Form (IPAQ-SF): a systematic review. Int. J. Behav. Nutr. Phys. Act 8, 115 (2011). [PubMed: 22018588]

72. Pei D, Jones CNO, Bhargava R, Chen Y-DI & Reaven GM Evaluation of octreotide to assess insulin-mediated glucose disposal by the insulin suppression test. Diabetologia 37, 843–845 (1994). [PubMed: 7988789]
73. Lam HYK et al. Detecting and annotating genetic variations using the HugeSeq pipeline. Nat. Biotechnol. 30, 226–229 (2012). [PubMed: 22398614]
74. Kalia SS et al. Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): a policy statement of the American College of Medical Genetics and Genomics. Genet. Med. 19, 249–255 (2017). [PubMed: 27854360]
75. Trapnell C, Pachter L & Salzberg SL TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25, 1105–1111 (2009). [PubMed: 19289445]
76. Love MI, Huber W & Anders S Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15, 1–21 (2014).
77. Contrepois K, Jiang L & Snyder M Optimized Analytical Procedures for the Untargeted Metabolomic Profiling of Human Urine and Plasma by Combining Hydrophilic Interaction (HILIC) and Reverse-Phase Liquid Chromatography (RPLC)-Mass Spectrometry. Mol. Cell. Proteomics 14, 1684–1695 (2015). [PubMed: 25787789]
78. Contrepois K et al. Cross-Platform Comparison of Untargeted and Targeted Lipidomics Approaches on Aging Mouse Plasma. Sci. Rep. 8, 17747 (2018). [PubMed: 30532037]
79. Lang RM et al. Recommendations for cardiac chamber quantification by echocardiography in adults: an update from the American Society of Echocardiography and the European Association of Cardiovascular Imaging. J. Am. Soc. Echocardiogr. 28, 1–39.e14 (2015). [PubMed: 25559473]
80. Wilson PWF et al. Prediction of Coronary Heart Disease Using Risk Factor Categories. Circulation 97, 1837–1847 (1998). [PubMed: 9603539]
81. Smith DA In adults without CVD, the MESA score, including coronary artery calcium, predicted 10-y risk for CHD events. Ann. Intern. Med. 164, JC35 (2016). [PubMed: 26974735]
82. McClelland RL et al. 10-Year Coronary Heart Disease Risk Prediction Using Coronary Artery Calcium and Traditional Risk Factors: Derivation in the MESA (Multi-Ethnic Study of Atherosclerosis) With Validation in the HNR (Heinz Nixdorf Recall) Study and the DHS (Dallas Heart Study). J. Am. Coll. Cardiol. 66, 1643–1653 (2015). [PubMed: 26449133]
83. Lee KK, Cipriano LE, Owens DK, Go AS & Hlatky MA Cost-effectiveness of using high-sensitivity C-reactive protein to identify intermediate- and low-cardiovascular-risk individuals for statin therapy. Circulation 122, 1478–1487 (2010). [PubMed: 20876434]
84. Myers J, Bader D, Madhavan R & Froelicher V Validation of a specific activity questionnaire to estimate exercise tolerance in patients referred for exercise testing. Am. Heart J. 142, 1041–1046 (2001). [PubMed: 11717610]
85. Arena R, Myers J, Aslam SS, Varughese EB & Peberdy MA Technical considerations related to the minute ventilation/carbon dioxide output slope in patients with heart failure. Chest 124, 720–727 (2003). [PubMed: 12907564]
86. Kaminsky LA, Imboden MT, Arena R & Myers J Reference Standards for Cardiorespiratory Fitness Measured With Cardiopulmonary Exercise Testing Using Cycle Ergometry: Data From the Fitness Registry and the Importance of Exercise National Database (FRIEND) Registry. Mayo Clin. Proc. 92, 228–233 (2017). [PubMed: 27938891]
87. Hovorka R, Soons PA & Young MA ISEC: a program to calculate insulin secretion. Comput. Methods Programs Biomed. 50, 253–264 (1996). [PubMed: 8894385]
88. Kamburov A, Cavill R, Ebbels TMD, Herwig R & Keun HC Integrated pathway-level analysis of transcriptomics and metabolomics data with IMPaLA. Bioinformatics 27, 2917–2918 (2011). [PubMed: 21893519]
89. Shannon P et al. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Res. 13, 2498–2504 (2003). [PubMed: 14597658]
90. Fruchterman TMJ & Reingold EM Graph drawing by force-directed placement. Softw. Pract. Exp. 21, 1129–1164 (1991).
91. Montagna PA Using SAS to Manage Biological Species Data and Calculate Diversity Indices. in 2014 SCSUG Educational Forum (, 2014).

92. Caporaso JG et al. QIIME allows analysis of high-throughput community sequencing data. Nat. Methods 7, 335–336 (2010). [PubMed: 20383131]

93. Callahan BJ et al. DADA2: High-resolution sample inference from Illumina amplicon data. Nat. Methods 13, 581–583 (2016). [PubMed: 27214047]

94. Bokulich NA et al. Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. Microbiome 6, 90 (2018). [PubMed: 29773078]

95. Callahan BJ, McMurdie PJ & Holmes SP Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. ISME J. 11, 2639 (2017). [PubMed: 28731476]

96. Chang CC et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 4, 7 (2015). [PubMed: 25722852]

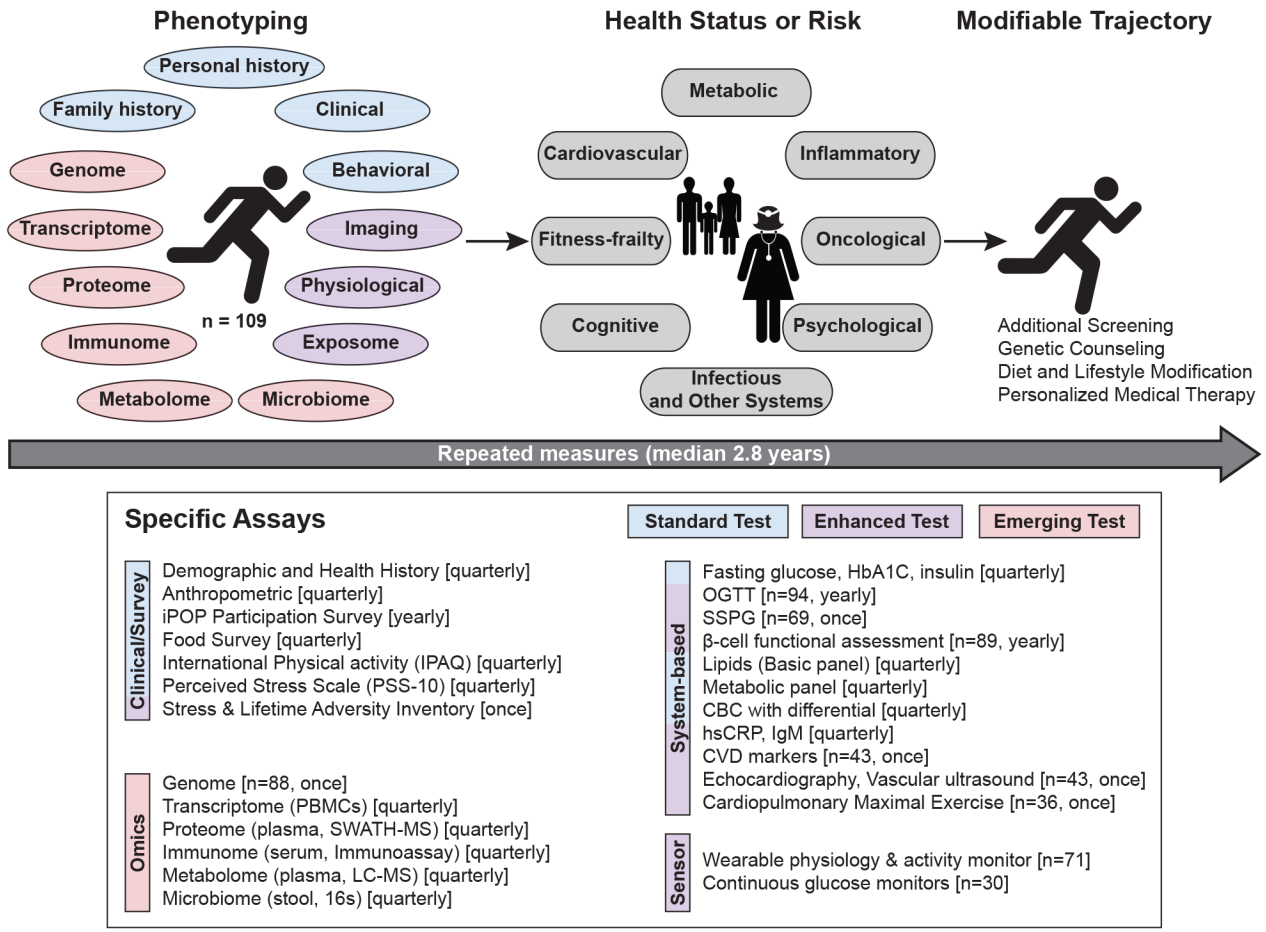


Figure 1. Study design and data collection.

Overview of the in-depth longitudinal phenotyping used to determine health risk and status. Data types were categorized as: Standard (Blue), Enhanced (Purple) and Emerging (Red) tests. PBMCs: peripheral blood mononuclear cells; HbA1C: glycated hemoglobin; OGTT: oral glucose tolerance test; SSPG: steady-state plasma glucose; CBC: complete blood count; hsCRP: high sensitivity C-reactive protein; CVD: cardiovascular disease.

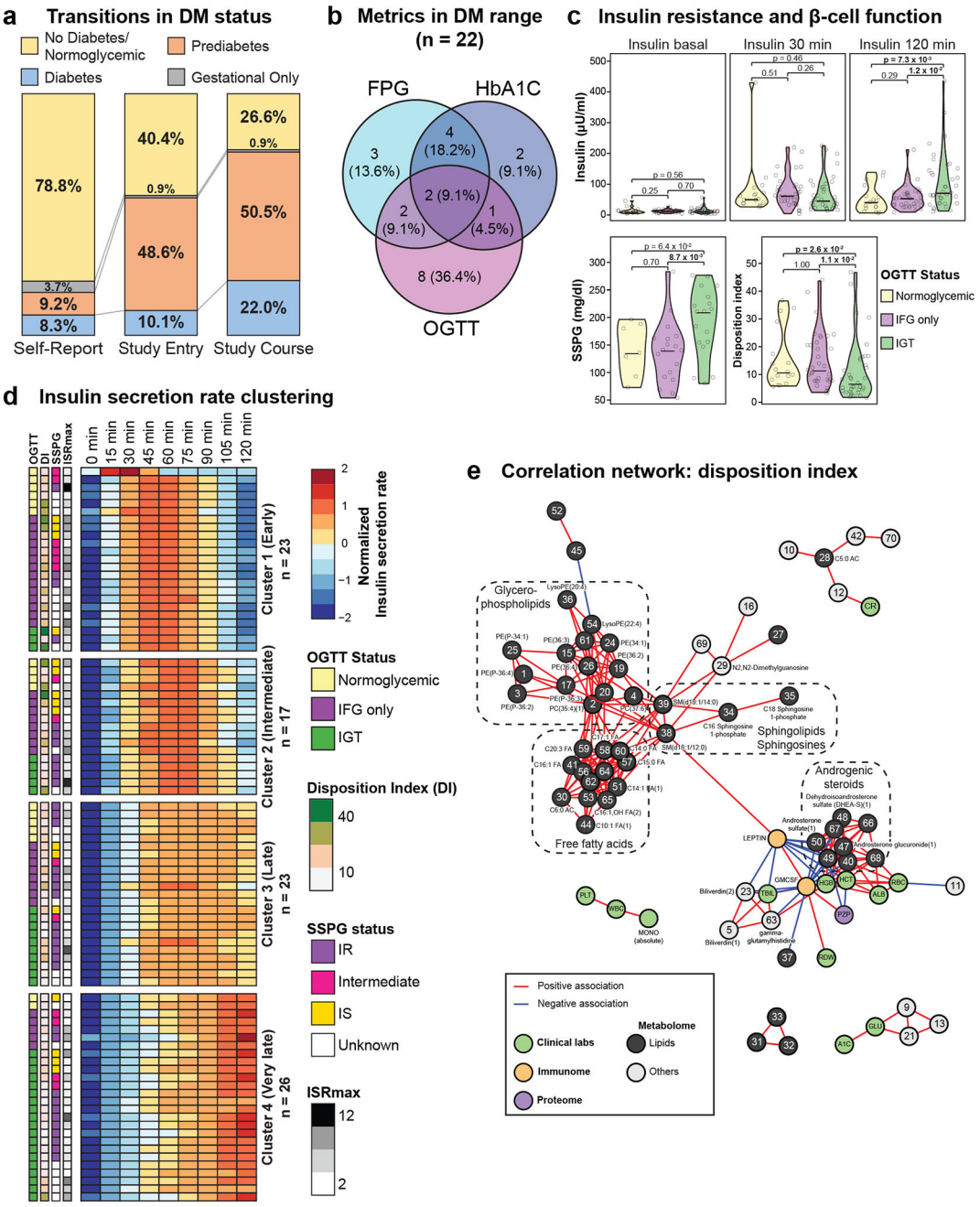


Figure 2. Clinical and enhanced phenotyping of glucose metabolism, insulin production and resistance.

(**a**) Transitions in diabetes mellitus (DM) status ($n = 109$). 1st column: Self-reported DM status; 2nd column: DM status determined by self-report; medical records and study entry diabetes-related laboratory measures: FPG, HbA1C and OGTT; prediabetic range (100 mg/dL ≤ FPG < 126 mg/dL or 5.7% ≤ HbA1C < 6.5% or 140 mg/dL ≤ OGTT < 200 mg/dL); diabetic range (FPG ≥ 126 mg/dL or HbA1C ≥ 6.5% or OGTT (2-hour) ≥ 200 mg/dL); 3rd column: DM history and status determined by the initial report and diabetes-related laboratory measures over the course of the study. For FPG to be considered impaired or

diabetic, two values in these ranges were required over the course of the study, whereas for HbA1C and OGTT only one value was required. (**b**) Overlap of diabetic range labs by participants over the course of the study. Diabetic ranges are as in panel (a). (**c**) Violin plots showing insulin levels during OGTT at 0, 30 and 120 minutes, SSPG (steady-state plasma glucose, $n = 43$ participants) and glucose disposition index ($n = 89$ samples from 61 participants) by glycemic status determined by OGTT including normoglycemic, impaired fasting glucose only (IFG only: FPG ≥ 100 mg/dL), and impaired glucose tolerance (IGT: OGTT ≥ 140 mg/dL). SSPG was measured using the modified insulin suppression test. The disposition index was calculated as the insulin secretion rate at 30 minutes times the Matsuda index (pmol/kg/min). A two-sided Wilcoxon t-test was used for differential analysis. The violin plots illustrate kernel probability density (*i.e.* the width represents the proportion of the data) and the horizontal bar depicts the median of the distribution. (**d**) Heatmap showing insulin secretion rates which were row-standardized and clustered using k-mean clustering ($n = 89$ samples from 61 participants). Observations within clusters were ordered by OGTT status. OGTT status, disposition index (DI), SSPG and insulin secretion rate max (ISR) are indicated on the left side of the heatmap. (**e**) Correlation network of multi-omics measures associated with the glucose disposition index (n = 89 samples from 61 participants; Benjamin-Hochberg FDR < 0.1). Correlations were calculated using Spearman correlation and considered significant if Bonferroni FDR < 0.05. Only networks containing a minimum of three molecules were plotted.

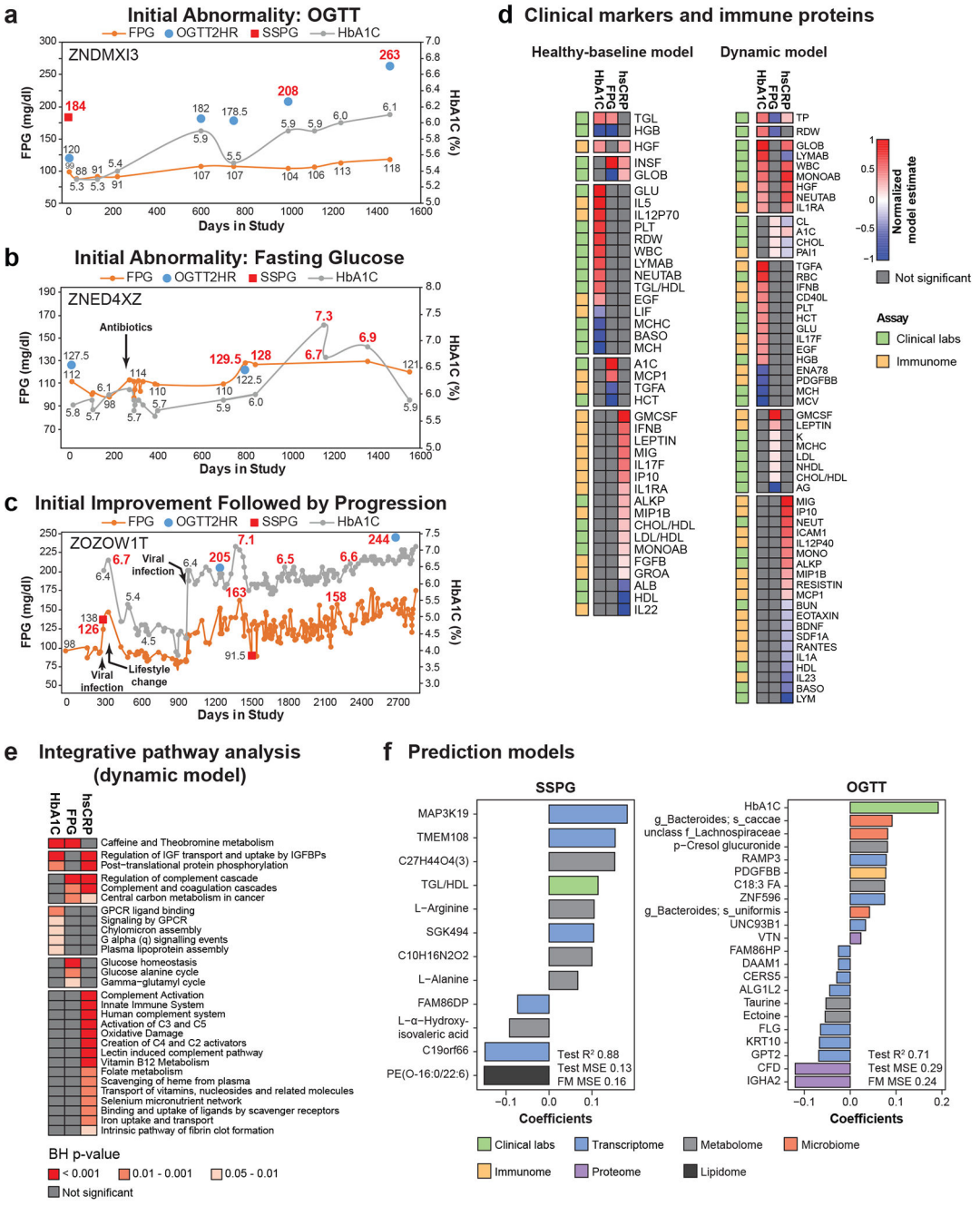


Figure 3. Longitudinal individual phenotyping and multi-omics of glucose metabolism and inflammation.

Longitudinal diabetic measures demonstrating different patterns of DM onset and progression with (**a**) initial abnormality response to glucose load (OGTT), (**b**) initial abnormality in fasting glucose metabolism (FPG) and (**c**) initial improvement followed by progression. Diabetic-range metrics are indicated in red. (**d**) Clinical markers and immune proteins associated with HbA1C, FPG, and hsCRP using healthy-baseline and dynamic models. Healthy-baseline models are linear mixed models that take into account repetitive measures across participants (HbA1C $n = 101$, samples 560; FPG $n = 101$, samples 563;

hsCRP n = 98, samples 518). Dynamic models are similar models except that analytes are normalized across individuals to the first measurement and all time points in the study are used (HbA1C n = 94, samples = 836; FPG n = 94, samples = 843; hsCRP n = 92, samples 777). Each analyte was modeled separately and the two sided t-test was used to determine p-value for each analyte effect. Multiple testing correction was performed and molecules were considered significant when Benjamin-Hochberg (BH) FDR < 0.2. Model estimates were normalized in each condition so the maximum value equal 1 and the minimal value equal −1. (**e**) Integrative pathway analysis using IMPaLa[66] of proteins and metabolites associated with HbA1C (n = 94, samples = 836), FPG (n = 94, samples = 843), and hsCRP (n = 92, samples 777) as determined by the dynamic models (BH FDR < 0.2 at molecule level). Significance of pathways was determined by the hypergeometric test (one-sided) followed by Fisher's combined probability test (one-sided) to determine combined pathway significance (BH FDR < 0.05). The n's of proteins and metabolites for each pathway are provided in Tables S15, S17 and S19. (**f**) Molecules selected in steady-state plasma glucose (SSPG) and oral glucose tolerance test (OGTT) prediction models and associated coefficients. For SSPG prediction, lipidomics data were used in addition to the multi-omics measures. MSE: mean square error.

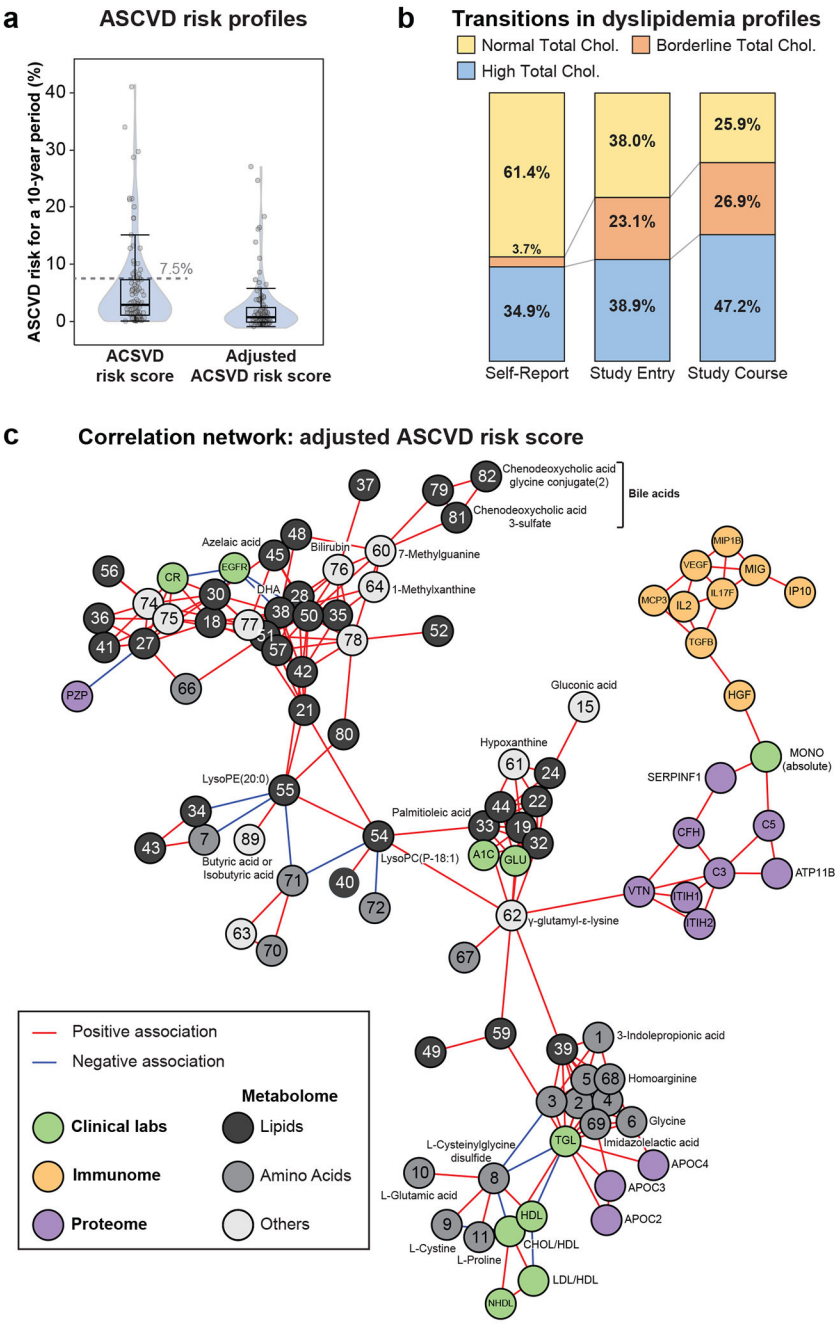


Figure 4. Clinical longitudinal cardiovascular health profiling and multi-omics correlation network of adjusted ASCVD risk.

(**a**) Distribution of ASCVD risk scores and adjusted ASCVD risk scores (n = 108). The box plot shows the 1st (lower edge of box), median (middle line) and 3rd (upper edge of box) quartiles. The upper whisker is the 3rd quartile + 1.5*(interquartile range) and the lower wisker is the lowest data point. (**b**) Self-reported cholesterol status versus measured total cholesterol profiles at study entry and over the course of the study (n = 108). (**c**) Multi-omics correlation network of molecules associated with adjusted ASCVD risk score (n = 77 participants) using Spearman correlation and multiple testing correction of q-value < 0.2.

Correlations between molecules were then calculated using Spearman correlation and considered significant if Bonferroni corrected p-value < 0.1. Only molecules belonging to the main network were plotted.

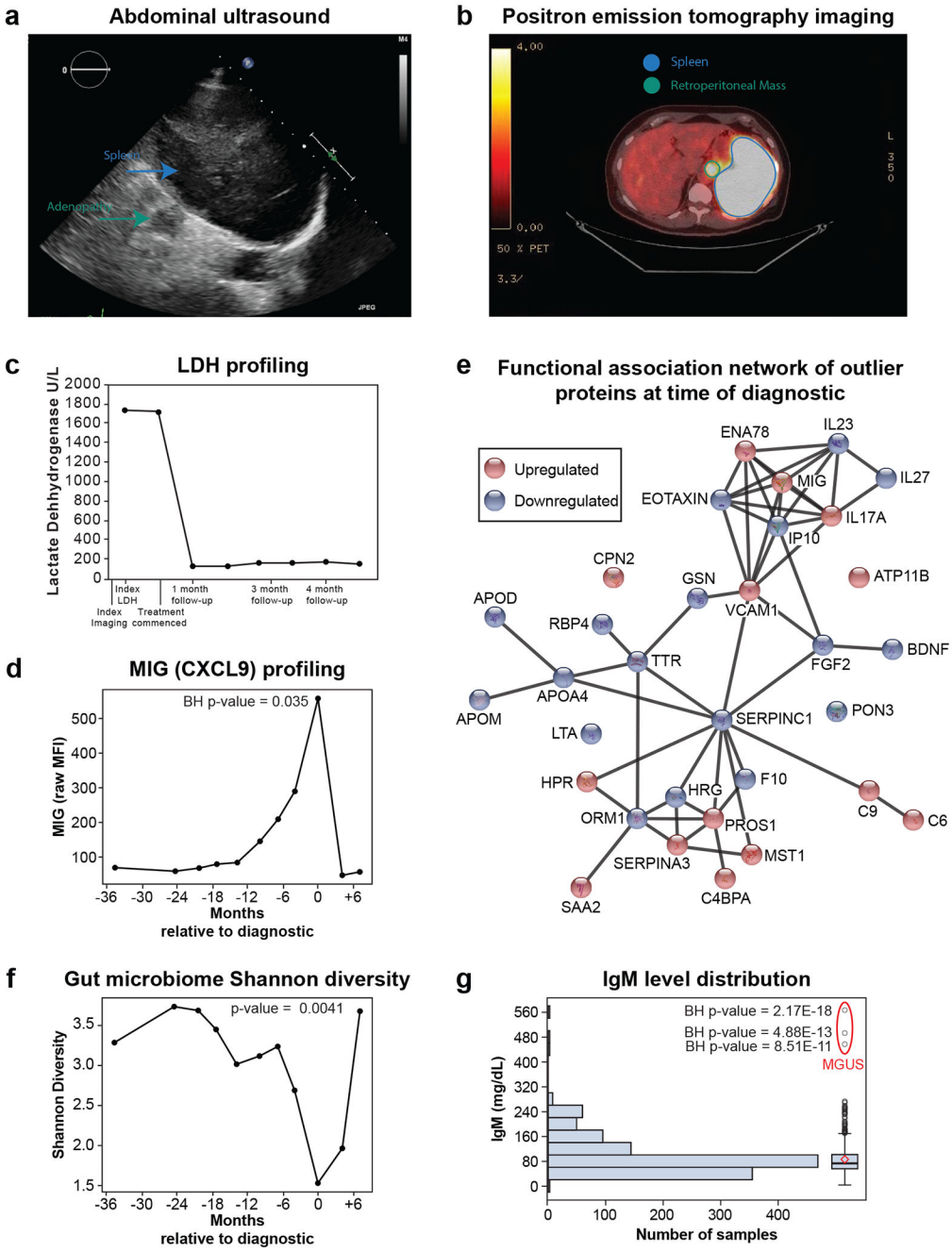


Figure 5. Oncologic discoveries.

(**a**) Abdominal ultrasound image where a mildly enlarged spleen measuring approximately 13 cm in craniocaudal dimension can be seen. (**b**) Positron emission tomography (PET) imaging where a large retroperitoneal mass with high fluorodeoxyglucose (FDG) and intensely focal hypermetabolism occupying the majority of the spleen can be seen. (**c**) Lactate Dehydrogenase (LDH) levels at time of index imaging and after starting chemotherapy. (**d**) Levels of MIG (CXCL9) demonstrating an increase starting a year prior to diagnosis that peaks at time of diagnosis and goes back to baseline after treatment (n=11 samples). Benjamin-Hochberg (BH) p-value (two-sided) was calculated on MIG Z-scores

assuming a normal distribution across all healthy visits in the cohort (n = 601 samples). (**e**) Functional association network of outlier proteins (95th percentile) at time of diagnostic. This analysis was performed using the web-tool STRING[67] (https://version-10-5.string-db.org/). Edges correspond to known, predicted or other interactions. (**f**) Shannon diversity of the gut microbiome decreasing months prior to diagnosis, reaching a minimum value at time of diagnostic and returning to baseline after treatment (n = 11 samples). Trajectory was then modeled using a general additive model which separates the linear ($\beta = -0.197$, p = 0.002 (2-sided t-test)) and non-linear (df = 3, p = 0.0112 (one-sided Chi-sq)) components. An F-test (one-sided) was used to compare the model including time to the null model. (**g**) IgM (Immunoglobulin M) level distribution in the cohort (n = 109, samples 1,111). Benjamin-Hochberg (BH) p-value (two-sided) was calculated on IgM Z-scores assuming a normal distribution across all visits in the cohort. Outlier visits are from a participant that was diagnosed with monoclonal gammopathy of undetermined significance (MGUS). The box plot shows the 1st (lower edge of box), median (middle line) and 3rd (upper edge of box) quartiles. The upper whisker is the 3rd quartile + 1.5*(interquartile range) and the lower wisker is the lowest data point. The diamond is the mean.

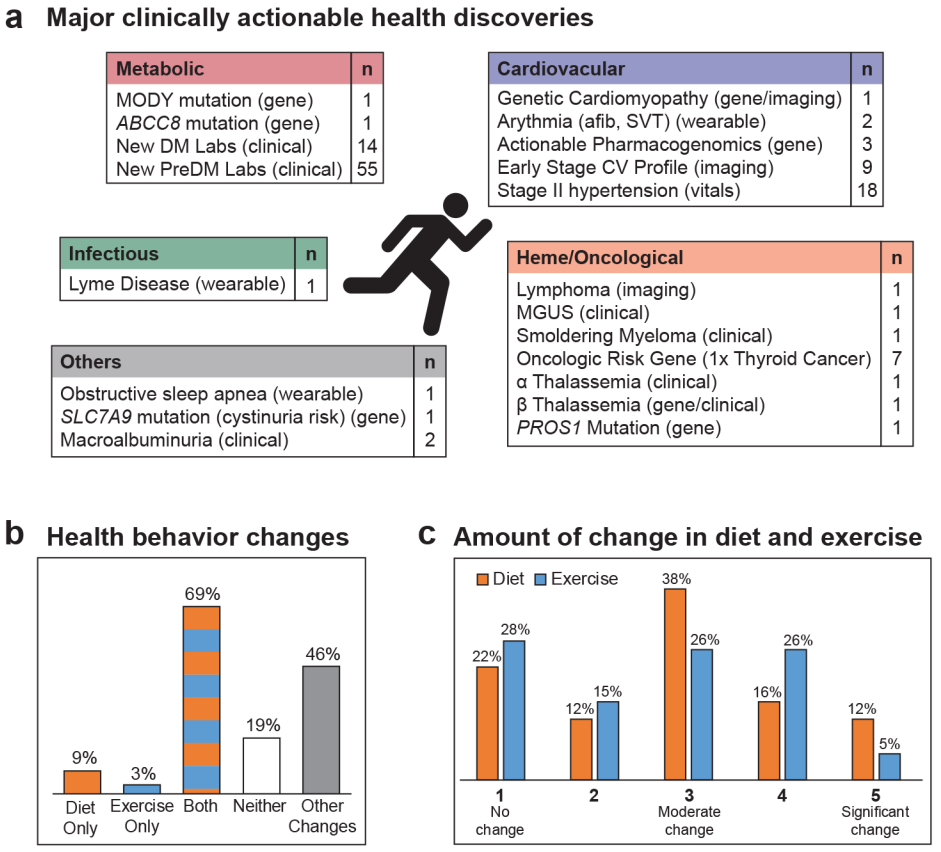


Figure 6. Summary of major clinically actionable health discoveries and participant health behavior change.

(**a**) Summary of clinically relevant health discoveries. 67 discoveries were considered major and the 55 PreDM results were not included in this count. (**b**) Diet and physical activity modifications. (**c**) Amount of change made in diet and exercise (5-point scale was used with 1 being no change and 5 being significant change). MODY: Maturity onset diabetes of the young; DM: diabetes mellitus; PreDM: prediabetes mellitus; afib: atrial fibrillation; SVT: supraventricular tachycardia; CV: cardiovascular; MGUS: monoclonal gammopathy of undetermined significance.